

# **Novel Statistical Methodology Development and Applications in ALS Research**

By  
© 2017  
Alex Karanevich  
M.S., University of Wyoming, 2014

Submitted to the graduate degree program in Biostatistics and the Graduate Faculty of the  
University of Kansas in partial fulfillment of the requirements for the degree of Doctor of  
Philosophy.

---

Co-Chair: Dr. Byron Gajewski

---

Co-Chair: Dr. Jianghua He

---

Dr. Devin Koestler

---

Dr. Jo Wick

---

Dr. Marjorie Bott

---

Dr. Jeffrey Statland

Date Defended: 13 November 2017

The dissertation committee for Alex Karanevich certifies that this is the  
approved version of the following dissertation:

**Novel Statistical Methodology Development and Applications in  
ALS Research**

---

Co-Chair: Dr. Byron Gajewski

---

Co-Chair: Dr. Jianghua He

Date Approved: 13 November 2017

## Abstract

Being able to predict, with accuracy, the disease progression of patients with a given disease is extremely useful from the perspectives of clinicians, patients, and clinical trial investigators. We introduce a novel method of reducing the expected prediction error when using linear models, given approximate monotonicity of the response; we refer to this method as utilizing an “anchor.” We justify this method mathematically, and then show how to improve predictions arising from standard ordinary least squares (OLS) models when modelling disease progression in a population of patients with amyotrophic lateral sclerosis (ALS).

We go on to show that using an anchor can be used in conjunction with more complex modelling schemes to further improve the predictions of ALS patients; an anchor improves both Bayesian hierarchical linear models and Bayesian mixture models. Furthermore, we explore potential covariates that may be included in the models to improve predictions, but find that only time of disease onset results in improved model performance.

We also explore how well these models work in a clinical setting, rather than in a clinical trial. We first demonstrate the feasibility of automatically extracting patients’ data, pertaining to survival and disease progression, from the electronic medical record, as well as showing that our disease progression model is feasible for clinical patients. We then compare survival rates between the two populations and determine that, even after adjusting for several important covariates, there is a large difference between survival in the clinic setting and survival in ALS clinical trials. We assert that the two patient groups’ differences in disease progression and survival highlight the needs to understand better disease variability in the clinical setting and to refine the inclusion criteria in ALS trials.

We determine an anchor can be used to improve predictive models in ALS disease progression, for both simple independent OLS regressions and for far more complicated Bayesian hierarchical linear models. We conclude that using a Bayesian hierarchical linear model with an anchor is useful in both a clinical trial population of ALS patients as well as a dissimilar population seen in the Midwestern academic medical center ALS clinic.

## Table of Contents

Novel Statistical Methodology Development and Applications in ALS Research .....	i
Abstract .....	iii
Table of Contents .....	v
Introduction .....	1
References .....	4
Chapter 1: Using an Anchor to Improve Linear Predictions .....	5
1.0 Abstract .....	5
1.1 Introduction .....	5
1.2 Theoretical Results for Simple Linear Regression .....	8
1.3 Application to ALS Prediction .....	18
1.4 Discussion .....	24
1.5 References .....	25
Chapter 2: Using an Onset-Anchored Bayesian Hierarchical Model to Improve Predictions for Amyotrophic Lateral Sclerosis .....	27
2.0 Abstract .....	27
2.1 Introduction .....	28
2.2 Methods .....	29
Study population. ....	29
Model comparison. ....	31
Bayesian hierarchical linear model. ....	32
Bayesian hierarchical linear mixture model.....	35

Bayesian onset-anchored hierarchical linear model.....	36
Covariate selection using the onset-anchored model.....	38
2.3 Results.....	40
2.4 Discussion.....	49
2.5 Conclusions.....	52
2.6 References.....	53
Chapter 3: Automated Data Extraction of Electronic Medical Records to Model ALS Survival	
and Progression.....	57
3.0 Abstract.....	57
3.1 Introduction.....	58
3.2 Methods .....	59
Study design.....	59
Study populations.....	60
Outcomes. ....	61
Statistical methods. ....	62
3.3 Results.....	64
Comparison of datasets. ....	64
Analysis of disease progression. ....	65
Analysis of survival. ....	66
3.4 Discussion.....	68
3.5 References.....	71
Summary and Future Directions .....	75
Appendix.....	77

Appendix A1: Full model description .....	77
Appendix A2: BUGS code (R2 OpenBUGS format) .....	78
Appendix A3: Full table the effects of covariates on prediction under the onset anchored model .....	78
Appendix B1: My Experience at the KUMC ALS Clinic .....	81
Appendix B2: ALSFRS and ALSFRS-R Questionnaire .....	83
Appendix B3: Funding and Support .....	87

## Introduction

Predictive modelling, while always an important aspect of statistics, recently has enjoyed a meteoric surge in popularity and attention due to the advent of big data and improved machine learning techniques. One interesting result of this is the rise of competitive data prediction challenges, such as the well-known Netflix challenge (1) which tasked teams with the goal of trying to produce an algorithm to determine users' ratings of given movies. Data prediction challenges occurred in all sorts of industrial and academic fields, including medicine (2). One challenge in particular, the ALS Stratification Challenge, tasked teams with trying to build predictive models that could predict the disease progression of patients with amyotrophic lateral sclerosis (ALS), also known as Lou Gehrig's disease (3).

ALS is a rarely occurring neurodegenerative disease that causes full-body paralysis and eventually death in those afflicted by it. Progression of the ALS is determined by the ALS Functional Rating Scale (ALSFRS) or the ALSFRS-Revised (ALSFRS-R), which is ten or twelve clinician-administered questions resulting in a score between zero and 40 (ALSFRS) or 48 (ALSFRS-R). A higher score corresponds to more function, and as the disease progresses (causing function to deteriorate) the ALSFRS score decreases, with scores under 24 representing serious loss of function.

Because ALS has large amounts of between-subject variability and is rarely occurring, ALS is a very difficult disease to study; trials require very large sample sizes to be adequately powered to accommodate the variability, which is difficult due to ALS's rarity. To assist in this, the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) database was created (4). PRO-ACT consists of data from dozens of ALS clinical trials (all individually negative, meaning



no significant improvement was found when using the treatment), and contains ALSFRS, lab values, survival information, and more for thousands of ALS patients.

Data in PRO-ACT was used by participants in the ALS Stratification challenge to both train and validate models. The winners of this challenge used random forest models (3), which are nonparametric and consequently are difficult to explain and interpret. This was the second large-scale crowd-sourced ALS data prediction challenge where the winners all utilized popular nonparametric approaches: the first being the DREAM Phil Bowen ALS Prediction challenge by Prize4life (5). The result of both of these challenges was that all top-performing teams used nonparametric models, such as Bayesian trees and random forests. Because of the linearity of the ALSFRS and ALSFSR-R, a parametric predictive linear model could also be a good candidate for measuring ALS progression. A linear model would be especially useful because of its easily-interpreted parameters.

There would be many potential advantages for having a successful explainable predictive model for ALS progression for not only ALS patients and their doctors, but also for those who are interested in developing new ALS therapeutics (5). One particular hurdle to developing such a model comes from the fact that many patients in the challenge had very few measurements: some patients had as little as one single data-point with which to construct a prediction. Because of this, even though the disease progression was fairly linear over time (6, 7), standard linear modelling techniques encountered difficulties; fitting a linear model through one or even two points allowed for too much variability in the predictions. This led us to develop a new methodology to reduce the variability of predictions, which we refer to as using an “anchor.” The anchor is an additional data point used in the regression that is created by intelligently assuming a particular value for the response when the process being modelled first began.

In chapter one of this dissertation we describe the mathematics which prove that including an anchor in the context of ordinary least squares (OLS) regression will reduce the variability of the resulting predictions. This comes at the cost of possibly increasing the bias of these predictions. We then go on to show how this method can be used in modelling disease progression of ALS patients by modelling each patient's disease progression independently. Since we have the actual disease progressions of each patient, we can easily compare the model that uses an anchor to one that does not, and see that the anchor results in dramatically improved predictive performance. Additionally, we see that the anchor allows us to make predictions much sooner and with less data than the traditional OLS model can.

In chapter two, we consider a more rigorous approach of modelling the disease progression of ALS patients. We formally test various linear models via cross-validation, ultimately deciding on a Bayesian hierarchical linear model using an anchor as the best model. We conduct a cross-validation to assess the benefit of including various covariates in said model. We conclude that the only covariate that improves prediction for a patient's disease progression is knowledge of when the patients first began having symptoms of ALS.

Finally, in chapter three we discuss the feasibility of creating disease progression and survival data models using ALS patient data that is automatically extracted from the electronic medical record (EMR). We look at the differences in disease progression and survival rates between ALS clinic patients and clinical trial subjects. We conclude that there is an urgent need to better understand disease variability in ALS patients, as well as a need to refine the inclusion criteria in ALS clinical trials.

## References

1. Greene K. The \$1 million Netflix challenge. Technology Review. [www.technologyreview.com/read\\_article.aspx](http://www.technologyreview.com/read_article.aspx). 2006 Oct.
2. Abdallah K, Hugh-Jones C, Norman T, Friend S, Stolovitzky G. The Prostate Cancer DREAM Challenge: A community-wide effort to use open clinical trial data for the quantitative prediction of outcomes in metastatic prostate cancer.
3. Zach N, Kueffner R, Atassi N, Chio A, Cudkowicz M, Hardiman O, Stolovitzky G. The ALS Stratification Prize-Using the Power of Big Data and Crowdsourcing for Catalyzing Breakthroughs in Amyotrophic Lateral Sclerosis (ALS)(P5. 102). *Neurology*. 2016 Apr 5;86(16 Supplement):P5-102.
4. Atassi N, Berry J, Shui A, Zach N, Sherman A, Sinani E, et al. The PRO-ACT database: design, initial analyses, and predictive features. *Neurology*. 2014;83(19):1719-25.
5. Küffner R, Zach N, Norel R, Hawe J, Schoenfeld D, Wang L, Li G, Fang L, Mackey L, Hardiman O, Cudkowicz M. Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nature biotechnology*. 2015 Jan 1;33(1):51-7.
6. Gordon PH, Cheng B, Salachas F, Pradat PF, Bruneteau G, Corcia P, et al. Progression in ALS is not linear but is curvilinear. *J Neurol*. 2010;257(10):1713-7.
7. Ikeda K, Hirayama T, Takazawa T, Kawabe K, Iwasaki Y. Relationships between disease progression and serum levels of lipid, urate, creatinine and ferritin in Japanese patients with amyotrophic lateral sclerosis: a cross-sectional study. *Intern Med*. 2012;51(12):1501-8.

## **Chapter 1: Using an Anchor to Improve Linear Predictions**

### **1.0 Abstract**

Linear models are some of the most straightforward and commonly used modelling approaches. Consider modelling approximately monotonic response data arising from a time-related process. If one has knowledge as to when the process began or ended, then one may be able to leverage additional assumed data to reduce prediction error. This assumed data, referred to as the “anchor,” are treated as an additional data-point generated at either the beginning or end of the process. The response value of the anchor is equal to an intelligently selected value of the response (such as the upper bound, lower bound, or 99<sup>th</sup> percentile of the response, as appropriate). The anchor reduces the variance of prediction at the cost of a possible increase in prediction bias, resulting in a potentially reduced overall mean-square prediction error. This can be extremely effective when few individual data points are available, allowing one to make linear predictions using as little as a single observed data point. We develop the mathematics showing the conditions under which an anchor can improve predictions, and demonstrate using this approach to reduce prediction error when modelling the disease progression of patients with amyotrophic lateral sclerosis.

### **1.1 Introduction**

Prediction always has been an important part of statistical modeling. With the advent of big data and the rise of machine learning, one may think that researchers have moved beyond prediction via simple linear models. This is not the case, however, especially in the field of medical research; a quick search of PubMed from January 2016 through July 2017 results in over 1000 publications that utilize linear (but not generalized linear) models. This is because linear models are usually one of the first attempted approaches when analyzing new data, and

surprisingly often, they are sufficient. Linear models are simple to calculate, requiring tiny amounts of computing power compared to some of the more complex machine-learning algorithms (such as neural networks). Most importantly, linear models are very straightforward to interpret and explain; a direct contrast to the more sophisticated “black-box” methods that are dependent on large datasets. The ability to interpret and understand statistical models, or model intelligibility, especially is important in the field of healthcare (1).

Yet linear models have their failings, especially when modelling a bounded response. Consider attempting to model the disease progression over time of a patient with amyotrophic lateral sclerosis (ALS), also known as Lou Gehrig’s disease. This is measured by the instrument known as the ALS Functional Rating Scale – Revised, or ALSFRS-R (2). The ALSFRS-R is always an integer between zero and 48, with 48 representing no spread of the disease and zero being the theoretical maximal spread of the disease. The progression of the ALSFRS-R tends to be very linear (3, 4), but because of its bounded nature, simple linear models have the inherent structural defect of creating predictions that violate the lower and upper bounds. Many adjustments to this problem exist; examples include truncating the prediction to 48 if the prediction is too large (0 if too small) (5) or performing a logistic transform on the data (6). If the goal is prediction (e.g., the patient’s ALSFRS-R at one year), these adjustments may not perform well when small amounts of observed data exist. The small number of data points can result in the variance of the prediction being very large, producing a large mean-squared prediction error (MSPE). Recall the MSPE is equivalent to the sum of the variance and squared bias of the prediction.

In this paper we consider a simple method to reduce the variability of linear predictions at the cost of potentially increasing the predictive bias. Biased linear regression itself is not new

(ridge regression (7) is one well-known example), but we do this in a unique way by exploiting our knowledge of when the process we are modeling (e.g. the patient's disease progression) first began.

Tracking the date when a patient first began noticing symptoms of ALS (their disease onset time) is common practice in ALS clinics and trials. From a modelling perspective, one could use this information in a variety of ways; the most obvious way is using it as a covariate in the model. Using a different approach, if we were to assume their ALSFRS-R score at roughly the time of their disease onset, what might their ALSFRS-R be? One could argue that the patient has had minimal, if any, disease progression at time of disease onset. It seems reasonable that one could assume their ALSFRS-R to be 48 (meaning the minimum possible disease progression) at this time. We could then create a new observation with ALSFRS-R score of 48 at the time of disease onset, and include that as one of the observations (data points) used to build our linear model.

In this paper we consider using knowledge of when a process starts to create an assumed data point that can be used to reduce variability of linear model predictions. We found no previous literature on this technique in our literature search. First we show how the inclusion of this point mathematically reduces the variance component of the MSPE under the assumptions of ordinary least-squares (OLS) linear regression. Then we calculate the bias component it brings to the MSPE; and we deduce the condition under which this approach can reduce the MSPE in predication combined variance and bias. Afterwards we give an example using this approach in the context of modeling ALS disease progression, showing how it improves the MSPE when compared to a linear model lacking the extra data point. We show how it is also superior to a logistic transform approach. We stress that this method is a simple to understand, easy to

perform, and inexpensive to implement approach. It is our hope that this approach may be used by pragmatic researchers to improve their linear predictions and estimations at very little additional cost.

## 1.2 Theoretical Results for Simple Linear Regression

Here we develop the theoretical results that justify the creation and use of an extra assumed data point to improve modelling. We shall refer to this data point as the “anchor.” Consider  $n - 1$  ordered pairs  $\{(x_i, y_i)\}$ ,  $i \in 1 \dots n - 1$ , where  $y_i$  is some response corresponding to  $x_i$ . As per ordinary linear regression (8), assume that  $x_i$  and  $y_i$  have a linear relationship, meaning that for some constants  $\beta_0$  and  $\beta_1$ ,  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ , with independent error terms  $\epsilon_i \sim N(0, \sigma^2)$ . Furthermore, assume an additional observation referred to as the “anchor” given by  $(x_n, y_n)$ , where  $y_n$  is some fixed constant in  $\mathbb{R}$ .

Consider the problem of predicting a new value  $y_0$  corresponding to a given  $x_0$ , that typically is obtained by using the OLS estimates for  $\beta_0$  and  $\beta_1$ , denoted as  $a$  and  $b$ . Denote the resultant prediction for  $y_0$  that uses the first  $n - 1$  coordinate pairs by  $\hat{y}_0^{(n-1)} = a^{(n-1)} + b^{(n-1)}x_0$ , and the prediction that also includes the anchor by  $\hat{y}_0^{(n)} = a^{(n)} + b^{(n)}x_0$ . Denote the errors between our prediction and the truth to be  $e_0^{(n-1)} = y_0 - \hat{y}_0^{(n-1)}$  and  $e_0^{(n)} = y_0 - \hat{y}_0^{(n)}$ . Recall that the variance of  $e_0^{(n-1)}$  (that was built from  $n - 1$  ordered pairs of data in standard OLS regression) is equivalent to:

$$\text{var}(e_0^{(n-1)}) = \text{var}(y_0 - \hat{y}_0^{(n-1)}) = \sigma^2 \left( 1 + \frac{1}{n-1} + \frac{(\bar{x}^{(n-1)} - x_0)^2}{\sum_{i=1}^{n-1} (x_i - \bar{x}^{(n-1)})^2} \right),$$

where  $\text{var}(e_0^{(n)}) = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(\bar{x}^{(n)} - x_0)^2}{\sum_{i=1}^n (x_i - \bar{x}^{(n)})^2} \right)$  represents the variance of the prediction error

obtained from utilizing all  $n$  data points (meaning we include the anchor).

We first show that  $\text{var}(e_0^{(n)}) \leq \text{var}(e_0^{(n-1)})$ , meaning any choice of anchor will decrease the variance component of the MSPE. We then derive an upper bound for the bias of the anchor such that the MSPE will decrease; in other words, how far away from the “true” line can the anchor be before it makes the MSPE worse.

Without loss of generality, we will assume the following for the observed data:

Assume that  $x_1, \dots, x_{n-1}$  have been normalized, meaning that  $\bar{x}^{(n-1)} = \frac{\sum_{i=1}^{n-1} x_i}{n-1} = 0$  and

$\sqrt{\sum_{i=1}^{n-1} x_i^2} = 1$ . Then the following hold:

$$\begin{aligned} SSX^{(n-1)} &= \sum_{i=1}^{n-1} x_i^2 = 1, \\ \bar{x}^{(n)} &= \frac{x_n}{n} + \frac{n-1}{n} (\bar{x}^{(n-1)}) = \frac{x_n}{n}, \\ SSX^{(n)} &= \sum_{i=1}^{n-1} x_i^2 + x_n^2 - n(\bar{x}^{(n)})^2 \\ &= 1 + x_n^2 - \frac{x_n^2}{n}. \end{aligned}$$

Any collection of  $(x_i, y_i)$  can be linearly transformed in the  $x$ -coordinate by subtracting the mean of the  $x$ 's and dividing by the Euclidean norm  $\sqrt{|| < x_1, \dots, x_{n-1} > ||}$  to achieve this. It is interesting to point out that this transformation has no impact on the OLS estimators for  $\sigma^2$ .

**Theorem 1:** For any anchor point  $(x_n, y_n)$ , with  $y_n$  a fixed constant,  $\text{var}(e_0^{(n)}) \leq \text{var}(e_0^{(n-1)})$ .

**Proof:** Let  $a, b$  be the OLS estimated intercept and slope through the points  $(x_1, y_1) \dots (x_n, y_n)$ .

In other words,  $a$  and  $b$  are the regression estimates for  $\beta_0$  and  $\beta_1$ . Since  $y_0$  and  $\hat{y}_0^{(n)}$  are independent,



$var(e_0^{(n)}) = var(y_0 - \hat{y}_0^{(n)}) = var(y_0) + var(a + bx_0)$ . Utilizing our assumptions on

$x_1, \dots, x_{n-1}$ , the inequality  $var(e_0^{(n)}) \leq var(e_0^{(n-1)})$  holds if and only if:

$$\begin{aligned} var(y_0) + var(a + bx_0) &\leq var(e_0^{(n-1)}) \\ &= \sigma^2 \left( 1 + \frac{1}{n-1} + \frac{(\bar{x}^{(n-1)} - x_0)^2}{SSX^{(n-1)}} \right) \\ &= \sigma^2 \left( 1 + \frac{1}{n-1} + \frac{(0 - x_0)^2}{1} \right). \end{aligned}$$

Which simplifies as follows:

$$\begin{aligned} \sigma^2 + var(a) + x_0^2 var(b) + 2x_0 cov(a, b) &\leq \sigma^2 \left( 1 + \frac{1}{n-1} + x_0^2 \right), \\ var(a) + x_0^2 var(b) + 2x_0 cov(a, b) &\leq \sigma^2 \left( \frac{1}{n-1} + x_0^2 \right). \end{aligned}$$

We next consider the individual terms  $var(a)$ ,  $var(b)$ , and  $cov(a, b)$ . For convenience  $SSX$  denotes  $SSX^{(n)}$  and  $\bar{x}$  denotes  $\bar{x}^{(n)}$ .

### Part 1: variance of slope

$$var(b) = var\left(\sum_{i=1}^n \frac{(x_i - \bar{x})}{SSX} y_i\right) = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{SSX^2} var(y_i).$$

Recall  $var(y_i) = \sigma^2$  if  $i \leq n-1$  and  $var(y_n) = 0$  since  $y_n$  is a constant. Thus:

$$var(b) = \frac{\sigma^2}{SSX^2} \sum_{i=1}^{n-1} (x_i - \bar{x})^2 = \frac{\sigma^2}{SSX^2} \sum_{i=1}^{n-1} (x_i^2 + \bar{x}^2 - 2x_i \bar{x}).$$

Utilizing the assumption that  $\sum_{i=1}^{n-1} (x_i^2) = 1$  and that  $\sum_{i=1}^{n-1} (x_i) = 0$ ,

$$var(b) = \frac{\sigma^2}{SSX^2} (1 + (n-1)\bar{x}^2).$$

Or equivalently

$$\text{var}(b) = \sigma^2 \frac{n^2 + nx_n^2 - x_n^2}{(nx_n^2 + n - x_n^2)^2}.$$

### Part 2: variance of intercept

Since  $\text{var}(y_n) = 0$ :

$$\begin{aligned} \text{var}(a) &= \text{var}\left(\sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SSX}\right) y_i\right) = \sigma^2 \sum_{i=1}^{n-1} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SSX}\right)^2 \\ &= \sigma^2 \sum_{i=1}^{n-1} \left(\frac{1}{n^2} + \frac{\bar{x}^2(x_i - \bar{x})^2}{SSX^2} - 2\frac{\bar{x}(x_i - \bar{x})}{nSSX}\right) \\ &= \sigma^2 \left(\frac{n-1}{n^2} + \frac{\bar{x}^2(1 + (n-1)\bar{x}^2)}{SSX^2} + 2\frac{(n-1)\bar{x}^2}{nSSX}\right). \end{aligned}$$

Which is equivalent to

$$\text{var}(a) = \sigma^2 \frac{nx_n^4 + 2nx_n^2 + n - x_n^4 - x_n^2 - 1}{(nx_n^2 + n - x_n^2)^2}.$$

### Part 3: covariance of intercept and slope

Consider  $\text{cov}(a, b)$ . We use the property that  $\text{cov}(\sum c_i y_i, \sum d_i y_i) = \sigma^2 \sum (c_i d_i)$  and the fact that any covariance or variance term involving  $y_n$  is 0, since  $y_n$  is a constant.

$$\begin{aligned} \text{cov}(a, b) &= \text{cov}\left(\sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SSX}\right) y_i, \sum_{i=1}^n \frac{(x_i - \bar{x})}{SSX} y_i\right) \\ &= \text{cov}\left(\sum_{i=1}^{n-1} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SSX}\right) y_i, \sum_{i=1}^{n-1} \frac{(x_i - \bar{x})}{SSX} y_i\right) \\ &= \sigma^2 \sum_{i=1}^{n-1} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SSX}\right) \left(\frac{(x_i - \bar{x})}{SSX}\right) \\ &= \frac{\sigma^2}{SSX} \sum_{i=1}^{n-1} \left(\frac{x_i - \bar{x}}{n} - \frac{\bar{x}(x_i - \bar{x})^2}{SSX}\right) \end{aligned}$$

$$= \frac{-\sigma^2}{SSX} \left( \frac{n-1}{n}(\bar{x}) + \frac{\bar{x}}{SSX} (1 + (n-1)\bar{x}^2) \right)$$

Or equivalently

$$\text{cov}(a, b) = -\sigma^2 \frac{x_n(nx_n^2 + 2n - x_n^2 - 1)}{(nx_n^2 + n - x_n^2)^2}.$$

**Part 4: Proving the inequality  $\text{var}_n(e_0) \leq \text{var}_{n-1}(e_0)$**

Recall,  $\text{var}(e_0^{(n)}) \leq \text{var}(e_0^{(n-1)})$  is equivalent to the following:

$$\text{var}(a) + x_0^2 \text{var}(b) + 2x_0 \text{cov}(a, b) \leq \sigma^2 \left( \frac{1}{n-1} + x_0^2 \right),$$

that is true if and only if

$$\begin{aligned} 0 \leq x_0^2 \left( 1 - \frac{n^2 + nx_n^2 - x_n^2}{(nx_n^2 + n - x_n^2)^2} \right) + x_0 \left( 2 \frac{x_n(nx_n^2 + 2n - x_n^2 - 1)}{(nx_n^2 + n - x_n^2)^2} \right) \\ + \left( \frac{1}{n-1} - \frac{nx_n^4 + 2nx_n^2 + n - x_n^4 - x_n^2 - 1}{(nx_n^2 + n - x_n^2)^2} \right). \end{aligned}$$

The right hand side of the inequality is quadratic in  $x_0^2$  with form  $g(x_0) = Ax_0^2 + Bx_0 + C$ . Note the coefficients  $A, B, C$  simplify in the following way:

$$A = \frac{(n-1)x_n^2(nx_n^2 + 2n - x_n^2 - 1)}{(nx_n^2 + n - x_n^2)^2},$$

$$B = \frac{2x_n(nx_n^2 + 2n - x_n^2 - 1)}{(nx_n^2 + n - x_n^2)^2},$$

$$C = \frac{(nx_n^2 + 2n - x_n^2 - 1)}{(n-1)(nx_n^2 + n - x_n^2)^2}.$$

Since  $A > 0$  for  $n > 2$ , then  $g(x_0)$  is an upward-facing parabola. Also, the discriminant, given by  $B^2 - 4AC$ , is equal to zero:

$$\begin{aligned}
B^2 - 4AC &= \frac{4x_n^2(nx_n^2 + 2n - x_n^2 - 1)^2}{(nx_n^2 + n - x_n^2)^4} \\
&\quad - 4 \frac{(n-1)x_n^2(nx_n^2 + 2n - x_n^2 - 1)}{(nx_n^2 + n - x_n^2)^2} \frac{(nx_n^2 + 2n - x_n^2 - 1)}{(n-1)(nx_n^2 + n - x_n^2)^2} \\
&= \frac{4x_n^2(nx_n^2 + 2n - x_n^2 - 1)^2}{(nx_n^2 + n - x_n^2)^4} - \frac{4x_n^2(nx_n^2 + 2n - x_n^2 - 1)^2}{(nx_n^2 + n - x_n^2)^4} \\
&= 0,
\end{aligned}$$

meaning there is exactly one root in  $g(x_0)$ . Therefore, it must be true that  $g(x_0) \geq 0$  and  $\text{var}(\hat{y}_0^{(n)}) \leq \text{var}(\hat{y}_0^{(n-1)})$  as desired. ■

Thus we see that any choice of anchor necessarily will result in a reduction in the variance of the prediction of  $y_0$ , that is equivalent to a reduction of the variance component of the MSPE. However, we still need to consider the bias. Recall that the typical OLS estimators for slope and intercept are unbiased. We next consider how much bias can be introduced before utilizing the anchor that no longer benefits the MSPE of our predictions. It will be shown that any choice of an anchor  $(x_n, y_n)$  such that  $y_n \neq \beta_0 + \beta_1 x_n$  will introduce bias to the model. Note that the bias is a direct function of  $\beta_0$  and  $\beta_1$ , that rarely are known in practice. Again, let  $\bar{x}$  denote  $\bar{x}^{(n)}$  and  $SSX$  denote  $SSX^{(n)} = \sum_{i=1}^n x_i^2$ .

**Theorem 2:** Using anchor point  $(x_n, y_n)$  results in biasing the slope by

$$E(b - \beta_1) = \frac{(n-1)x_n(y_n - \beta_0) + \beta_1 n}{nSSX} - \beta_1.$$

**Proof:** Recall  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$  and that the OLS estimate for  $\beta_1$ , denoted by  $b$ , is given by

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{SSX} = \frac{\sum_{i=1}^n \left(x_i - \frac{x_n}{n}\right)y_i}{SSX}.$$

We first derive  $E(b)$ :

$$\begin{aligned} E(b) &= E\left(\frac{\sum_{i=1}^n \left(x_i - \frac{x_n}{n}\right) y_i}{SSX}\right) \\ &= E\left(\frac{\sum_{i=1}^{n-1} \left(x_i - \frac{x_n}{n}\right) y_i}{SSX} + \frac{\left(x_n - \frac{x_n}{n}\right) y_n}{SSX}\right). \end{aligned}$$

Recall that  $y_n$  is a nonrandom constant, and hence  $E(y_n) = y_n$ . Then:

$$\begin{aligned} E(b) &= \frac{1}{SSX} E\left(\sum_{i=1}^{n-1} \left(x_i y_i - \frac{y_i x_n}{n}\right)\right) + \frac{1}{SSX} \left(x_n - \frac{x_n}{n}\right) y_n \\ &= \frac{1}{SSX} E\left(\sum_{i=1}^{n-1} \{x_i y_i\} - \bar{y}_{n-1} x_n \left(\frac{n-1}{n}\right)\right) + \frac{1}{SSX} \left(x_n - \frac{x_n}{n}\right) y_n. \end{aligned}$$

Note that the OLS estimate for  $\beta_0$  when not using the anchor point is given by

$$a^{(n-1)} = \bar{y}^{(n-1)} - b^{(n-1)} \bar{x}^{(n-1)} = \bar{y}^{(n-1)} \text{ since } \bar{x}^{(n-1)} = 0. \text{ Similarly, } b^{(n-1)} = \sum_{i=1}^{n-1} \{x_i y_i\}.$$

Since these are the unbiased OLS estimators for  $\beta_0$  and  $\beta_1$  when ignoring the anchor point, then

it must be that  $E(\bar{y}^{(n-1)}) = \beta_0$  and  $E(\sum_{i=1}^{n-1} \{x_i y_i\}) = \beta_1$ . Then we have

$$\begin{aligned} E(b) &= \frac{1}{SSX} \left(\beta_1 - \beta_0 x_n \left(\frac{n-1}{n}\right)\right) + \frac{1}{SSX} \left(x_n - \frac{x_n}{n}\right) y_n \\ &= \frac{1}{SSX} \left(\beta_1 + x_n y_n - \beta_0 x_n \left(\frac{n-1}{n}\right) - \frac{x_n y_n}{n}\right) \\ &= \frac{1}{SSX} \left(\beta_1 + \left(\frac{n-1}{n}\right) x_n y_n - \beta_0 x_n \left(\frac{n-1}{n}\right)\right) \\ &= \frac{1}{SSX} \left(\beta_1 + x_n \left(\frac{n-1}{n}\right) (y_n - \beta_0)\right). \end{aligned}$$

Or equivalently

$$E(b) = \frac{(n-1)x_n(y_n - \beta_0) + \beta_1 n}{nSSX},$$

that means the bias  $b$  is given by

$$E(b - \beta_1) = \frac{(n-1)x_n(y_n - \beta_0) + \beta_1 n}{nSSX} - \beta_1.$$

■

**Theorem 3:** Using anchor point  $(x_n, y_n)$  results in biasing the intercept by

$$E(a - \beta_0) = \frac{\beta_0(n-1)(x_n^2 + 1) - \beta_1 x_n^2 + y_n}{nSSX} - \beta_0.$$

**Proof:** Recall that the OLS estimate for  $\beta_0$ , denoted by  $a$ , is given by

$$a = \bar{y}^{(n)} - b\bar{x} = \bar{y}^{(n)} - \frac{bx_n}{n}.$$

We first calculate  $E(a)$ :

$$\begin{aligned} E(a) &= E(\bar{y}^{(n)}) - \frac{x_n}{n} E(b) \\ &= \frac{1}{n} E((n-1)\bar{y}^{(n-1)} + y_n) - \frac{x_n}{n} E(b). \end{aligned}$$

Again, recall that  $E(\bar{y}^{(n-1)}) = \beta_0$  and that  $E(y_n) = y_n$ . We derived  $E(b)$  in Theorem 2. Thus:

$$E(a) = \frac{(n-1)}{n} \beta_0 + \frac{y_n}{n} - \frac{x_n}{n^2 SSX} ((n-1)x_n(y_n - \beta_0) + \beta_1 n),$$

that can be reduced to

$$E(a) = \frac{\beta_0(n-1)(x_n^2 + 1) - \beta_1 x_n + y_n}{nSSX}.$$

Therefore the bias of the intercept is

$$E(a - \beta_0) = \frac{\beta_0(n-1)(x_n^2 + 1) - \beta_1 x_n + y_n}{nSSX} - \beta_0.$$

■

**Corollary:** The overall bias induced by using anchor point  $(x_n, y_n)$  is given by

$$\begin{aligned}
E(\hat{y}_0 - y_0) &= E(a + bx_0 - \beta_0 - \beta_1 x_0) \\
&= \frac{\beta_0(n-1)(x_n^2 + 1) - \beta_1 x_n + y_n}{nSSX} - \beta_0 + x_0 \left\{ \frac{(n-1)x_n(y_n - \beta_0) + \beta_1 n}{nSSX} - \beta_1 \right\}.
\end{aligned}$$

**Corollary:** The anchor point introduces no bias, meaning  $\hat{y}_0^{(n)}$  is unbiased, only if  $y_n = \beta_0 + \beta_1 x_n$ .

**Proof:** Recall from Theorem 2 we have

$$E(b) = \frac{1}{SSX} \left( \beta_1 - \beta_0 x_n \left( \frac{n-1}{n} \right) \right) + \frac{1}{SSX} \left( x_n - \frac{x_n}{n} \right) y_n.$$

Then if  $y_n = \beta_0 + \beta_1 x_n$  the following holds:

$$\begin{aligned}
E(b) &= \frac{1}{SSX} \left( \beta_1 - \beta_0 x_n \left( \frac{n-1}{n} \right) + x_n \left( \frac{n-1}{n} \right) (\beta_0 + \beta_1 x_n) \right) \\
&= \frac{1}{SSX} \left( \beta_1 + \beta_1 \frac{x_n^2(n-1)}{n} \right) = \frac{1}{SSX} \beta_1 \left( 1 + \frac{x_n^2(n-1)}{n} \right) = \beta_1 \left( \frac{SSX}{SSX} \right) = \beta_1.
\end{aligned}$$

Thus  $b$  is unbiased when  $y_n = \beta_0 + \beta_1 x_n$ . Recall from Theorem 3 that

$$\begin{aligned}
E(a) &= \frac{1}{n} E \left( (n-1) \bar{y}^{(n-1)} + y_n \right) - \frac{x_n}{n} E(b) \\
&= \frac{1}{n} ((n-1)\beta_0 + \beta_0 + \beta_1 x_n) - \frac{x_n}{n} \beta_1.
\end{aligned}$$

Using the derivation above for  $E(b) = \beta_1$ . This then simplifies to be

$$E(a) = \beta_0,$$

meaning  $a, b$  are unbiased and therefore  $E(\hat{y}_0^{(n)}) = E(a + bx_0) = \beta_0 + \beta_1 x_0 = y_0$ . ■

**Theorem 4:** Utilizing anchor point  $(x_n, y_n)$  reduces the overall MSPE when the following inequality holds:

$$\left( \frac{\beta_0(n-1)(x_n^2 + 1) - \beta_1 x_n + y_n}{nSSX} - \beta_0 + x_0 \left\{ \frac{(n-1)x_n(y_n - \beta_0) + \beta_1 n}{nSSX} - \beta_1 \right\} \right)^2 \leq$$

$$x_0^2 \left( \frac{(n-1)x_n^2(nx_n^2 + 2n - x_n^2 - 1)}{n^2SSX^2} \right) + x_0 \left( \frac{2x_n(nx_n^2 + 2n - x_n^2 - 1)}{n^2SSX^2} \right)$$

$$+ \frac{(nx_n^2 + 2n - x_n^2 - 1)}{(n-1)n^2SSX^2}.$$

Proof: Consider the following inequality

$$MSPE^{(n)} \leq MSE^{(n-1)}.$$

This is equivalent to

$$bias^2(e_0^{(n)}) + var(e_0^{(n)}) \leq bias^2(e_0^{(n-1)}) + var(e_0^{(n-1)})$$

$$bias^2(e_0^{(n)}) \leq var(e_0^{(n-1)}) - var(e_0^{(n)}).$$

But recall  $bias^2(e_0^{(n-1)}) = 0$ , and that  $var(e_0^{(n-1)}) = \sigma^2 \left( 1 + \frac{1}{n-1} + x_0^2 \right)$ . The remaining pieces,  $bias^2(e_0^{(n)})$  and  $var(e_0^{(n)})$ , were derived in Theorem 2 and Theorem 3, and substituting them in to this inequality results in the formula given in the statement of Theorem 4.

■

Thus, we see that any choice of anchor point will reduce the variance of prediction, but will increase the bias of the prediction depending on how far away the anchor point is from the “true” regression line. Therefore, using an anchor may be beneficial or not, depending on how much bias is added.

The bound calculated in Theorem 4 potentially could be used as a decision rule for determining if using an anchor is beneficial or not. Unfortunately, one needs to know the true values of  $\beta_0$  and  $\beta_1$  in order to use Theorem 4’s result. In practice, one tends not to know the true



regression parameters, although with sufficiently informed prior knowledge, precise estimates may exist. Thus, when deciding whether to use an anchor or not, we suggest comparing the anchor model to a standard model using a cross-validation approach. We show an example of this in the results section.

Before moving to the application section, we note that many of the ideas in this paper have Bayesian connections. For example, consider performing a Bayesian analysis of classical regression. When using the standard noninformative prior distribution, the posterior mean estimates for the slope and intercept terms (and their standard errors) are equivalent to those obtained under frequentist OLS (9). It follows that Theorems 1-4 still hold under the Bayesian paradigm, meaning that an anchor can be used to reduce the variance of posterior predictions.

### **1.3 Application to ALS Prediction**

We next consider using an anchor to improve linear models that pertain to predicting disease progression in patients with ALS. Note that the theory developed in part (2) applies to a single OLS regression (prediction for the individual). The following example expands on this by showing how using an anchor can improve the average prediction error across several OLS regressions (prediction for each of several individuals).

Our data come from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) database (10). In 2011, Prize4Life, in collaboration with the Northeast ALS Consortium, and with funding from the ALS Therapy Alliance, formed the PRO-ACT Consortium. The data available in the PRO-ACT Database have been volunteered by PRO-ACT Consortium members (<https://nctu.partners.org/PRO-ACT/>).

Recall ALS disease progression is tracked by the ALSFRS-R, our outcome variable, that is an integer value between zero and 48, where 48 represents the minimal amount of disease

progression and zero represents the maximal progression. For each patient, we model the ALSFRS-R versus time (in days). Specifically, time is measured in days from trial baseline, meaning  $x = 0$  corresponds to the beginning of the trial and  $x = 365$  corresponds to the 365<sup>th</sup> day after the trial began. On this scale, a patient's disease onset time is typically negative, as it happened before the trial began. We required patients to have the following: (1) at least two recorded ALSFRS-R scores before 3 months, for model building purposes; (2) non-missing value for time of disease onset; (3) at least one year between the baseline and last ALSFRS-R score for MSPE-validation purposes. This resulted in 1,606 patients, with an average  $\pm$  standard error (SE) of  $12 \pm 4.54$  time-points per patient (and  $3 \pm 0.96$  visits in the first three months).

Note that we are now considering longitudinal data. Volumes have been written on modelling strategies for data such as this; hierarchical models, mixed models, time series, and machine learning algorithms all would make good candidates for the modeling of this. However, to demonstrate how using the anchor-point improves OLS regression, we simply will model each patient independently with: (1) a standard OLS regression model and (2) with an OLS regression model using an anchor. Note that the ALSFRS-R follows a fairly linear decline, although each patient exhibits wide variation in their patient-specific progression rates, justifying using linear models (Figure 1.1). The assumed data point, or anchor, used in the anchor model comes from assuming minimal disease progression at the time of disease onset. In other words, each patient's data is augmented with the additional data point given by the ordered pair  $(x_{onset}, 48)$ , since 48 is the ALSFRS-R corresponding to minimal progression.

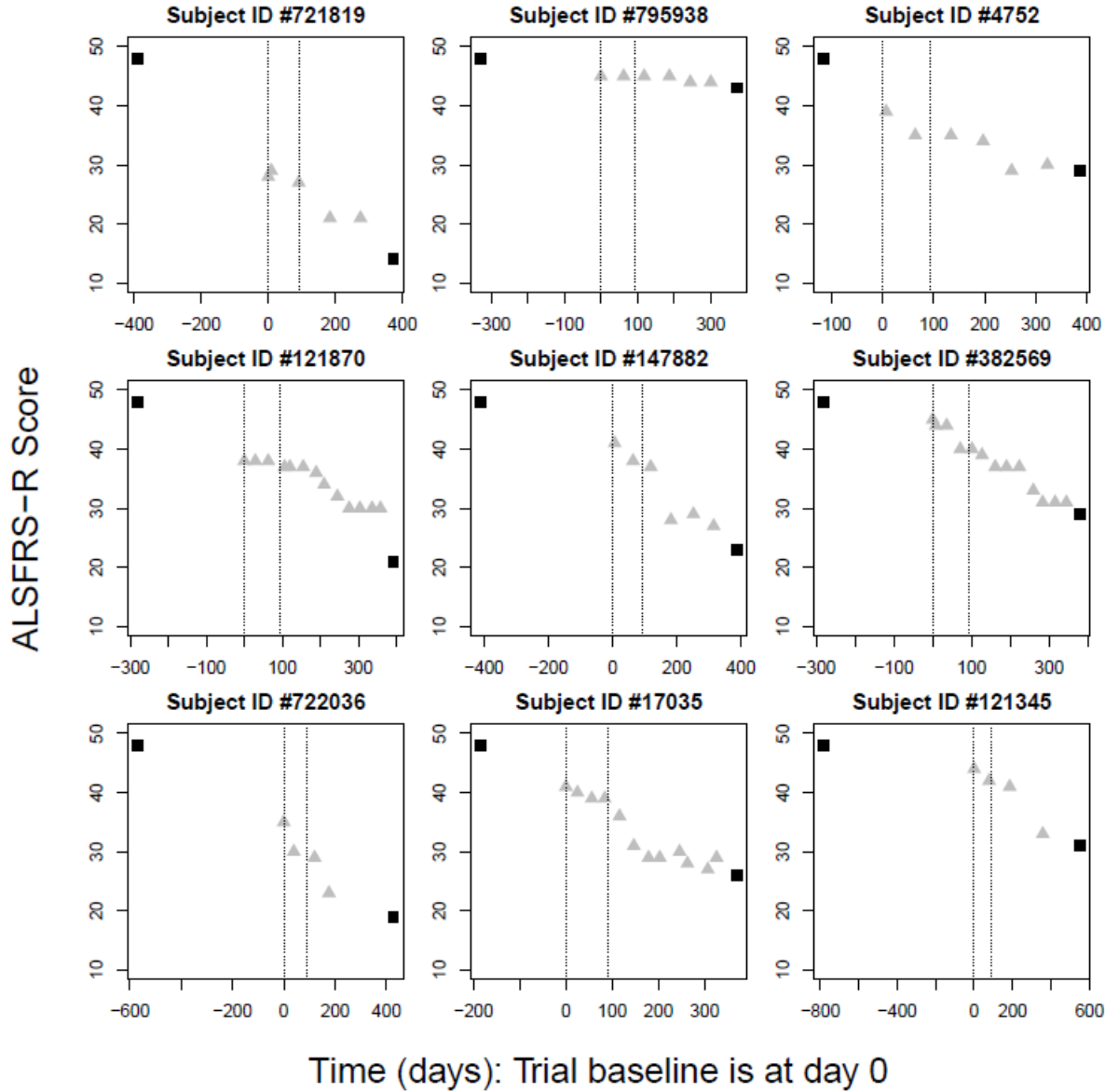


Figure 1.1: For nine randomly selected subjects, we plotted their ALSFRS-R versus time. The leftmost black square is the anchor, the rightmost square is the true value  $Y_k$ , and the gray triangles are observed scores. The dashed black lines denote days 0 and 92 of the trial, meaning observations between the two dashed lines were used for model fitting.

We will compare the standard model versus the anchor model by comparing their ability to predict each patient  $k$ 's first ALSFRS-R score after 365 days (1 year), observed at time  $x_{k,0}$ , using only ALSFRS-R scores measured before 92 days (3 months). Specifically for both models we calculate

$$\sqrt{MSPE} = \sqrt{\frac{\sum_{k=1}^{1606} (\hat{Y}_k - Y_k)^2}{1606}},$$

where  $\hat{Y}_k$  is the predicted ALSFRS-R score for patient  $k$  at time  $x_{k,0}$  and  $Y_k$  is the true ALSFRS-R score at time  $x_{k,0}$ . Because we know the ALSFRS-R is bounded between 0 and 48, any model prediction that falls outside these bounds will be truncated to the closest boundary value before evaluating the MSPE. To assist in visualizing this data, Figure 1.1 shows the progression of the ALSFRS-R versus time for nine subjects (simple random sample without replacement).

The anchor model results in slightly more biased predictions compared to those of the standard model, as expected. However, as demonstrated in the methods section, the variance of these errors is much smaller for the anchor model (Figure 1.2). The resulting root MSPE of the anchor model is 7.8 while the standard model's MSPE is 13.0; we observe a large drop in prediction error when including the anchor.

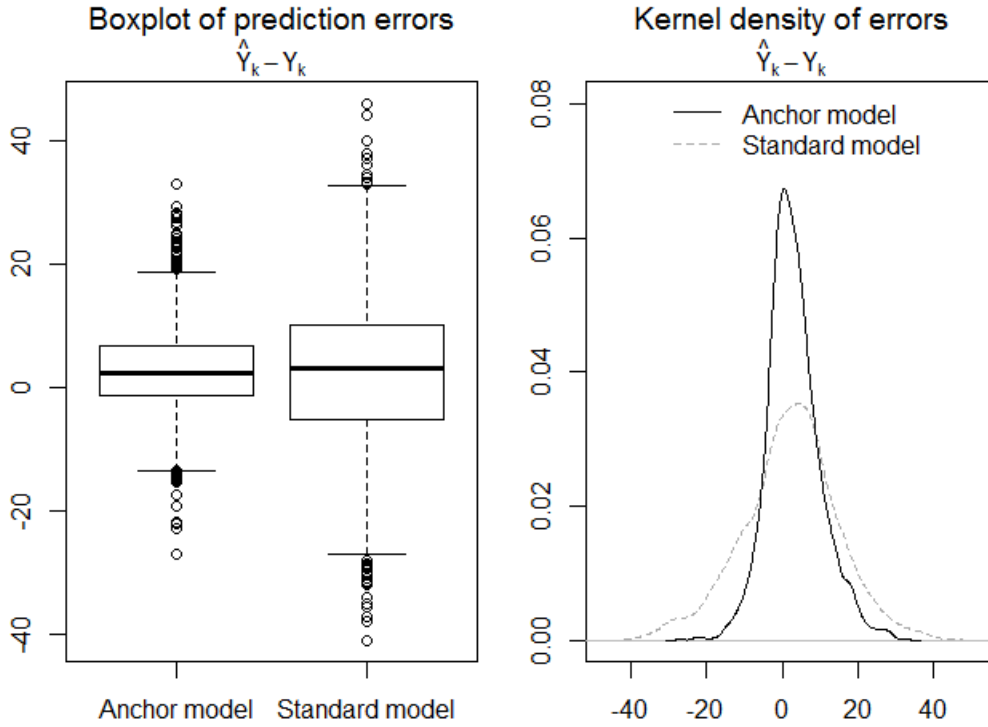


Figure 1.2: The raw prediction error for the anchor and standard models. The models' mean error as measured by  $\hat{Y}_k - Y_k$  was 3.1 and 2.1 respectively, with standard deviations of 7.1 and 12.7.

It can be shown that for some patients, the prediction from the standard model is closer to the truth than the prediction from the anchor model. Perhaps we should only use the anchor model when the increase in bias is negligible. We could explore this by taking the difference between the prediction from the anchor model  $\hat{Y}^{(a)}$  and the standard model  $\hat{Y}^{(s)}$ ; if this difference is sufficiently small in magnitude then the increase in bias from using the anchor model is negligible on average. In other words, for each patient consider calculating  $T_k = \hat{Y}_k^{(a)} - \hat{Y}_k^{(s)}$ , and then defining the prediction for patient  $k$  as

$$\hat{Y}_k = \begin{cases} \hat{Y}_k^{(a)} & \text{if } |T_k| \leq \Gamma \\ \hat{Y}_k^{(s)} & \text{otherwise} \end{cases}$$

for some constant  $\Gamma$ . Figure 1.3 shows how this changes the MSPE for various values of  $\Gamma$ , as well the result from changing the rule to be  $\hat{Y}_k = \hat{Y}_k^{(a)}$  if  $|T_k| \geq \Gamma$  instead (meaning choose the anchor model if the difference in the model predictions is large). From Figure 1.3 we see that naively using the anchor model for all patients outperforms any of the  $\Gamma$  and  $T_k$  decision-rule hybrids for this dataset.

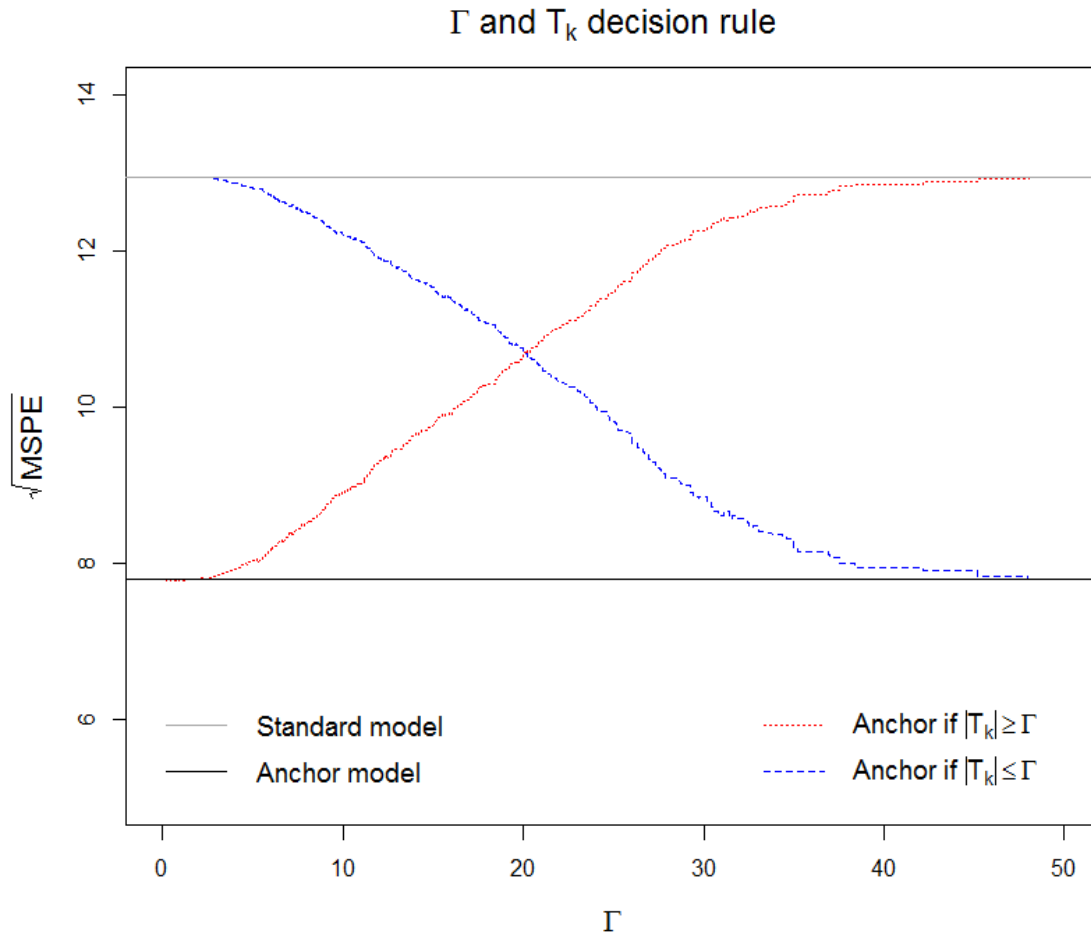


Figure 1.3: Shows the resulting MSPE for various cutoffs of  $\Gamma$ . Note that since the MSPE is bounded below by the anchor model ( $\sqrt{MSPE} = 7.78$ ), this shows that the anchor model is uniformly better than the linear model ( $\sqrt{MSPE} = 12.95$ ) for this data.

Finally, we compare the anchor model to that of a logistic transform model. The logistic transform is a model that is more advanced yet also more difficult to calculate and interpret. We

fit the logistic transform model by taking each ALSFRS-R score, dividing it by its maximum of 48, and fitting the resultant data (that is bounded between 0 and 1) with a logistic regression model. In other words, for a given patient we fit the following model:  $\text{logit}\left(\frac{y_i}{48}\right) = \beta_0 + \beta_1 x_i + \epsilon_{ij}$ , where  $\epsilon_{ij}$  are independent errors that follow  $N(0, \sigma^2)$ ,  $\beta_0$  and  $\beta_1$  are the intercept and slope parameters,  $x_i$  is the time-point associated with ALSFRS-R score  $y_i$ . The MSPE of this model comes to be 14.65, significantly higher than the MSPE for either the standard OLS model or the anchor model.

#### 1.4 Discussion

In this paper, we discussed a simple and computationally inexpensive technique that may improve the predictive power in linear models. This method consists of creating an additional assumed data point, referred to as an anchor, and including it in the OLS regression. It has been shown in this paper that including an anchor theoretically decreases prediction variance at the cost of potentially increased bias. We demonstrated how using an anchor can improve linear predictions from modelling disease progression in ALS patients.

Fitting the anchor model can be performed as easily and efficiently as a standard OLS regression, yet it has the potential to be a much stronger predictive model. Furthermore, the interpretations of the anchor model's parameters remain largely unchanged from that of OLS regression, which is a huge advantage over other models. The interpretability of the parameters is arguably one of the most attractive parts of linear models.

We imagine that using an anchor in the way we have demonstrated will be of particular use when modelling a bounded linear process where one can obtain a measure of when the process first began and/or ended. The bounds give a justification for choosing the y-value of the anchor; without bounds it may be difficult to justify a value without first looking at the data,

potentially leading to overfitting. However, as long as monotonicity approximately holds, one could still use something such as the 99<sup>th</sup> percentile of the response if no bound exists.

While the results in this paper are done under the assumptions of frequentist OLS regression, it is in no way limited to this. The idea of using this additional data point easily can extend to other families of models such as generalized linear models, hierarchical models, and mixed models. For example, one dramatically can improve the model performance in the ALS example by switching from independent linear regressions for each patient to a Bayesian hierarchical model; this allows patients to borrow information from one another and results in improved estimators due to shrinkage (11). This model is improved even further when it becomes a Bayesian hierarchical model that uses an anchor for each patient (12).

Deciding when to include an anchor for modelling is not straightforward. If the goal is estimation, the induced bias may not be worth the reduced variability in estimates. While we developed a theoretical bound for when an anchor will improve the MSPE, it depends on having theoretical knowledge of the underlying linear process, which is rarely possible in practice. Thus, we recommend using cross-validation to compare using an anchor versus a more standard approach, as we performed in our ALS example. Because some sort of cross-validation is good standard practice when evaluating predictive models, we feel that this is a very small price to pay for a potentially dramatic improvement in predictive error.

## 1.5 References

1. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Sydney, NSW, Australia. 2788613: ACM; 2015. p. 1721-30.



2. Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). J Neurol Sci. 1999;169(1-2):13-21.
3. Armon C, Graves MC, Moses D, Forte DK, Sepulveda L, Darby SM, et al. Linear estimates of disease progression predict survival in patients with amyotrophic lateral sclerosis. Muscle Nerve. 2000;23(6):874-82.
4. Magnus T, Beck M, Giess R, Puls I, Naumann M, Toyka KV. Disease progression in amyotrophic lateral sclerosis: predictors of survival. Muscle Nerve. 2002;25(5):709-14.
5. Amemiya T. Regression Analysis when the Dependent Variable Is Truncated Normal. Econometrica. 1973;41(6):997-1016.
6. Lesaffre E, Rizopoulos D, Tsonaka R. The logistic transform for bounded outcome scores. Biostatistics. 2007;8(1):72-85.
7. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 2000;42(1):80-6.
8. Kutner MH, Nachtsheim C, Neter J. Applied linear regression models. 4th ed. Boston ; New York: McGraw-Hill/Irwin; 2004. 701 p. p.
9. Gelman A. Bayesian data analysis. Third edition. ed. Boca Raton: CRC Press; 2014.
10. Atassi N, Berry J, Shui A, Zach N, Sherman A, Sinani E, et al. The PRO-ACT database: design, initial analyses, and predictive features. Neurology. 2014;83(19):1719-25.
11. Morris CN, Lysy M. Shrinkage Estimation in Multilevel Normal Models. 2012:115-34.
12. Karanevich AG, Statland JM, Gajewski BJ, He J. Using an onset-anchored Bayesian hierarchical model to improve predictions for amyotrophic lateral sclerosis disease progression (*In Review*). Manuscript submitted for publication. 2017.

## Chapter 2: Using an Onset-Anchored Bayesian Hierarchical Model to Improve Predictions for Amyotrophic Lateral Sclerosis

### 2.0 Abstract

**Background:** Amyotrophic Lateral Sclerosis (ALS), also known as Lou Gehrig’s disease, is a rare disease with extreme between-subject variability, especially with respect to rate of disease progression. This makes modelling a subject’s disease progression, which is measured by the ALS Functional Rating Scale (ALSFRS), very difficult. Consider the problem of predicting a subject’s ALSFRS score at 9 or 12 months after a given time-point.

**Methods:** We obtained ALS subject data from the Pooled Resource Open-Access ALS Clinical Trials Database, a collection of data from various ALS clinical trials. Due to the typical linearity of the ALSFRS, we consider several Bayesian hierarchical linear models. These include a mixture model (to account for the two potential classes of “fast” and “slow” ALS progressors) as well as an onset-anchored model, in which an additional artificial data point, using time of disease onset, is used to improve predictive performance.

**Results:** The onset-anchored model had a drastically reduced predictive mean-square-error, when compared to the Bayesian hierarchical linear model or the mixture model under a cross-validation approach. No covariates, other than time of disease onset, consistently improved predictive performance in either the Bayesian hierarchical linear model or the onset-anchored model.

**Conclusions:** Augmenting patient data with an additional artificial data point, or onset anchor, can drastically improve predictive modelling in ALS by reducing the variability of estimated parameters at the cost of a slight increase in bias. This onset-anchored model is extremely useful if predictions are desired directly after a single baseline measure (such as at the first day of a

clinical trial), a feat that would be very difficult without the onset anchor. This approach could be useful in modelling other diseases that have bounded progression scales (e.g. Parkinson's disease, Huntington's disease, or inclusion-body myositis). It is our hope that this model can be used by clinicians and statisticians to improve the efficacy of clinical trials and aid in finding treatments for ALS.

## 2.1 Introduction

Amyotrophic Lateral Sclerosis (ALS) is a rare neuro-degenerative disease that exhibits extreme between-subject variability. Progression of ALS typically is measured by the ALS Functional Rating Scale (known as the ALSFRS), or with additional respiratory questions (the revised ALSFRS-R). The ALSFRS is a physician-reported outcome on a scale of 0 – 40 that grades common activities of daily living like dressing, eating, and walking. An ALSFRS score of 40 corresponds to normal function, and this score will decrease as the disease progresses. The ALSFRS, that is usually non-increasing, has been shown to decrease in a linear fashion over the course of a typical clinical trial (6 months to 1 year) (1, 2), although the linearity is disputed over long periods of time (3).

Faster disease progression consistently is associated with lowered survival (2, 4-8). Many of the clinical measurements shown to be associated with survival (e.g. region of symptom onset and Riluzole use) are not significantly associated with disease progression (9-11). Riluzole is the only FDA-approved drug for ALS. As rates of progression on the ALSFRS often are used in phase II and III clinical trials, more accurate predictive models would help researchers in improving trial efficiency.

Our aim was to develop a predictive Bayesian hierarchical model that could be used to predict individual ALSFRS scores at one year from trial beginning using at most the first three

months of clinical trial data. Our baseline model is a Bayesian hierarchical linear model, which is similar to a linear mixed effects model. We then compared the predictive power of this baseline model to those provided by a Bayesian mixture model and a Bayesian onset-anchored hierarchical linear model. The onset-anchored model leverages an additional data point for each patient, which assumes maximum ALSFRS score at the time of disease onset. Note that the approach of using an onset-anchor is applicable in modelling other diseases that use a bounded rating scale (Parkinson's disease, Huntington's disease, etc.). We additionally consider variable selection to improve model predictive accuracy, as well as consider model robustness when less than three months of data are available.

## 2.2 Methods

**Study population.** The datasets analyzed during this study are available in the Pooled Resource Open-Access ALS Clinical Trials database (PRO-ACT) (<https://nctu.partners.org/ProACT/>) (12). In 2011, Prize4Life, in collaboration with the Northeast ALS Consortium, and with funding from the ALS Therapy Alliance, formed the PRO-ACT Consortium. The data available in the PRO-ACT Database have been volunteered by PRO-ACT Consortium members. As of December 2015, PRO-ACT had 4,838 unique subjects, each having at least one reported ALSFRS or ALSFRS-R score. As PRO-ACT is a collection of data from clinical trials, we further subset this data only to include subjects that were receiving placebos. This resulted in 1,301 subjects to be considered for analysis. One patient was dropped later due to having no data entered for disease onset time, bringing the final number of subjects to 1,300. For more demographic information on these subjects, see Table 2.1.

**Categorical Counts**

<b>Sex</b>	Male: 812	Female: 488	
<b>Race</b>	White: 1218	Black: 22	Latino: 13
	Asian: 12	Indian: 1	Other: 34
<b>Riluzole Use</b>	Yes: 600	No: 358	Not Reported: 342

<b>Continuous</b>	<b>Mean</b>	<b>SD</b>
<b>Age (at trial start)</b>	55.5	11.9
<b>Onset Time (days from trial start)</b>	-658.4	456
<b>Number of ALSFRS scores</b>	9.3	4.5

Table 2.1: Demographic data for  $n=1,300$  ALS subjects from the PRO-ACT database considered for analysis.

For these 1,300 subjects, we used ALSFRS scores to measure disease progression. The ALSFRS score is bounded between zero and 40, and typically is non-increasing. Patients with ALSFRS-R scores, the revised ALSFRS, had their scores converted to the ALSFRS by summing the scores from the first nine questions of the ALSFRS-R (that concern motor and bulbar function) as well as the score from the first respiratory question, R1: Dyspnea.

**Model comparison.** Our objective was to build a predictive model with which we could use the first three months of a subject's data to determine their ALSFRS score at one year. As very few subjects had a measurement at exactly one year, we instead used the model to predict each subject's first score after day 365, denoted as  $FRS_{365}$ . Three months was chosen as the cutoff because: 1) this was the window used in the DREAM ALS Stratification Prize4Life Challenge; 2) three months represented a reasonable amount of time for making 12-month predictions; and 3) this was a time frame with utility for both adaptive trial designs and for imputing missing data. Ideally, this model would be accurate even when less than three months of subject data are available.

Large amounts of variability are inherently associated with any ALS model. Bayesian hierarchical models excel at capturing many sources of variability that can be reported via posterior predictive credible intervals. A credible interval is preferable for its interpretability; in the framework of a Bayesian model, there is a 95% chance that a subject's  $FRS_{365}$  is within the 95% credible interval. Note to control the type I error rate, the gold standard for confidence intervals is 90% or 95%. A credible interval, being a statement of probability, has no such restriction, and thus is useful with even lower credible levels, such as 70% or 80% (13).

We considered three types of models that are described below: A Bayesian hierarchical linear model, a Bayesian hierarchical mixture model, and a Bayesian onset-anchored hierarchical linear model. Note that these models are all linear with respect to time. This is largely because a patient in PRO-ACT typically has only one ALSFRS measurement per month, which causes more complicated models, such as 3-parameter sigmoidal curves, to suffer from convergence problems. Linearity also is convenient because the slope parameter can be used as a simple-to-interpret measure of the disease's rate of progression.

The models were compared using their posterior mean-square-error (MSE) that resulted from a cross-validation analysis. Cross-validation entails splitting the data into a training set with which to build the model, and a validation set with which to assess the model's predictive power (13). The MSE is defined as the sum of the squared difference between the predicted  $FRS_{365}$  and the observed  $FRS_{365}$  for each subject in the validation set, and is adjusted for the size of the validation set.

In order to be in the validation set, subjects needed at least one ALSFRS score after one year from baseline. Again, as the ALSFRS score at one year specifically was not observed for most patients, we instead predicted  $FRS_{365}$ , the subject's first score after 365 days. Of subjects who had at least one year of data, average  $FRS_{365}$  was 386.7 days, with standard deviation of 23.7 days and maximum of 577 days. The same training and validation sets were used to validate all three models.

All analyses were done using R (14), OpenBUGS (15), and the R package R2openBUGS (16). Pseudo-code that describes the model in more detail is provided in the appendix (A2).

**Bayesian hierarchical linear model.** Since ALS seems to progress linearly over most one year time frames in the PRO-ACT database, we started with a linear hierarchical Bayes mixed effects model with weak and uninformative priors. Specifically, the ALSFRS for subject  $i$  at time  $t$  is modeled as:

$$ALSFRS_i(t) \sim T_3(a_i + b_i t, \sigma^2)$$

restricted to  $ALSFRS_i(t) \in [0,40]$ , that easily is done in OpenBUGS.  $T_3$  denotes the centered non-standardized t-distribution with three degrees of freedom and non-standardized variance  $\sigma^2$ . Note that a standardized t-distribution with three degrees of freedom would instead have a variance of one. Parameters  $a_i$  and  $b_i$  are the subject-specific intercept and slope term. A t-

distribution with three degrees of freedom was chosen because a normal distribution was too narrow in the tails. Additionally, we observed that the residuals from simple linear regression on subjects (with sufficient amounts of data) followed a  $T_3$  distribution extremely well (see Figure 2.1).

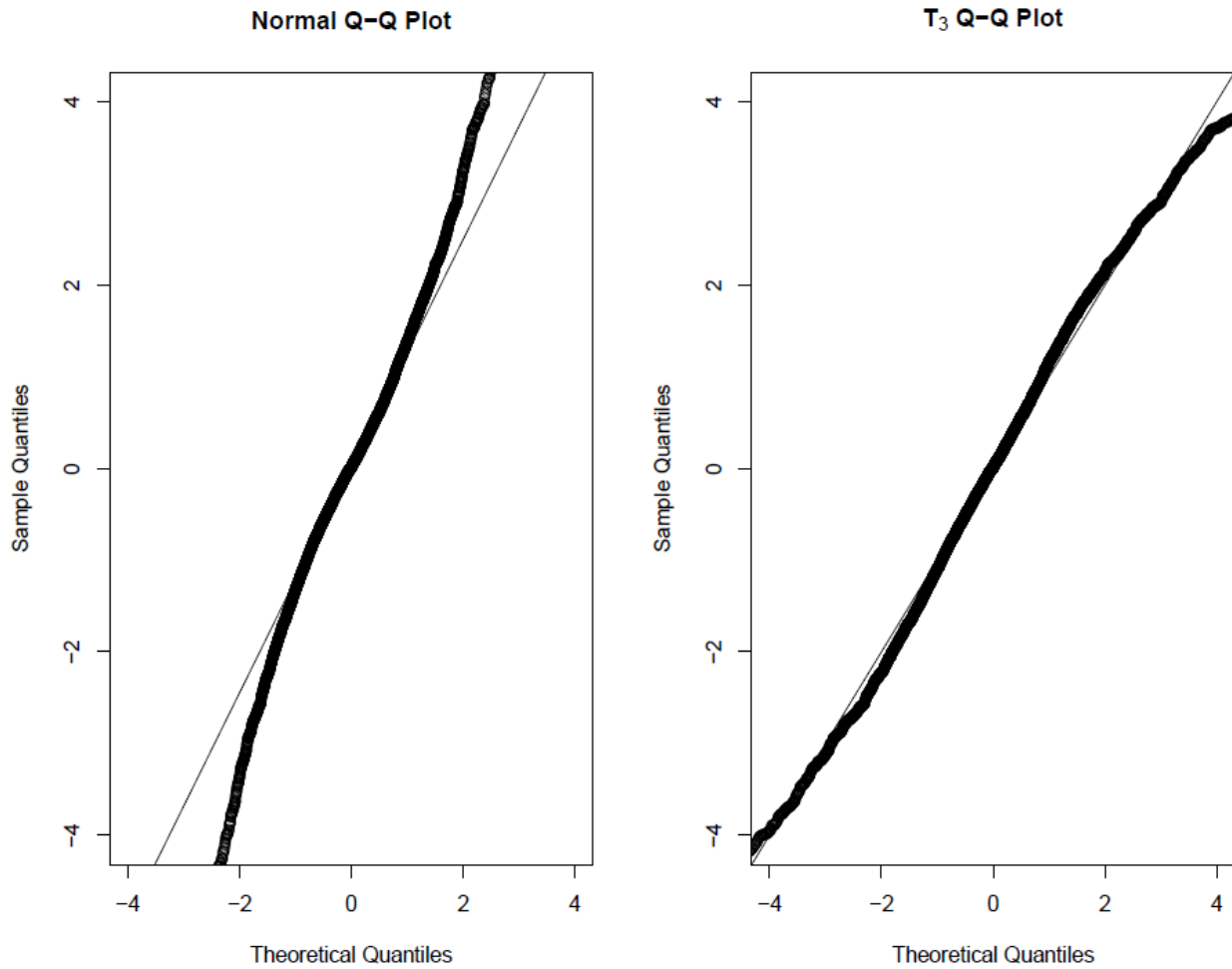


Figure 2.1: QQ plots for the residuals obtained from fitting simple linear regression models on subjects that had at least 5 ALSFRS measurements.

Continuing the model description, the hyper-parameters  $a_i$  and  $b_i$ , in turn, have the following distributions:

$$a_i \sim N(p_0, \sigma_0^2)$$

$$b_i \sim N(p_1, \sigma_1^2)$$



Where  $b_i$  is restricted to be non-positive. Weak priors from the literature and discussions with clinicians were assumed for  $p_0$  and  $p_1$ . Specifically, Castrillo-Viguera et. al. (17) reported that the ALSFRS-R decline in one database is roughly -0.92 units per month with a standard error of 0.08. This translates to roughly an ALSFRS decline of -.025 per day, and leads us to the following priors:

$$p_0 \sim N(33, 3^2)$$

$$p_1 \sim N(-0.025, 0.3^2)$$

Where the increased error in  $p_1$  allows for more strength in the analysis to come from the data. Generally subjects with low baseline ALSFRS scores are not enrolled in clinical trials, and the prior for  $p_0$  was chosen to reflect this while still allowing a wide range of potential starting ALSFRS values. Uninformative priors were assigned to the remaining variables.

Such a Bayesian model, aside from the weakly informed priors on  $p_i$ , was suggested by Gomeni et al. (18). A key advantage to hierarchical modelling in this way is that it allows for shrinkage of error resulting from sample means (19, 20), and also lets subjects with fewer data points “borrow” information from the remaining population. The Bayesian analysis also has advantages with respect to interpretability (especially in a clinical setting). This model will be referred to as the “linear model”.

**Bayesian hierarchical linear mixture model.** A mixture model is useful when each subject belongs to one of several groups, each group having their own specific progression distributions. Specifically, Gomeni et al. (18), suggested that ALS subjects could be classified as either “fast” or “slow” progressors. To model this, we assume each subject is either a fast or slow progressor, and assume that each group has their own average rate of disease progression (parameterized by the mean of the subject-specific slope). We further assume the slope parameter for fast progressors strictly is steeper (more negative) than those of slow progressors.

The ALSFRS for subject  $i$  at time  $t$  is still  $ALSFRS_i(t) \sim T_3(a_i + b_i t, \sigma^2)$ , but now we let  $b_i \sim N(\Lambda, \sigma_1^2)$ . This starts the mixture process, with  $\Lambda$  being either  $\Lambda_1$  or  $\Lambda_2 = (\Lambda_1 + c)$ , where  $c$  is a positive constant, with probability  $\Pr(\Lambda = \Lambda_i) = \pi_i$ . Finally, we use the following priors,  $\pi_i \sim \text{Dirichlet}(1,1)$ ,  $\Lambda_1 \sim N(0, \sigma_{\Lambda_1})$ . The remaining parameters from the mixture, such as  $c$  and  $\sigma_{\Lambda_1}$ , are assigned uninformative positive priors. All other priors and parameters are specified as in the linear model (2.2.1). This model shall be referred to as the “mixture model”.

**Bayesian onset-anchored hierarchical linear model.** This model resembles the linear model in structure, but uses an idea first introduced by Proudfoot et al. (21). The idea was to create an additional *artificial* data-point, referred to as the “onset-anchor”. We do this by assuming that each subject had an ALSFRS score of 40 (the maximum possible score) at their time of disease onset (see Figure 2.2). Aside from this artificial data point, the parameters and model specification remain identical to those given in the linear model. This model is referred to as the “onset-anchored model”.

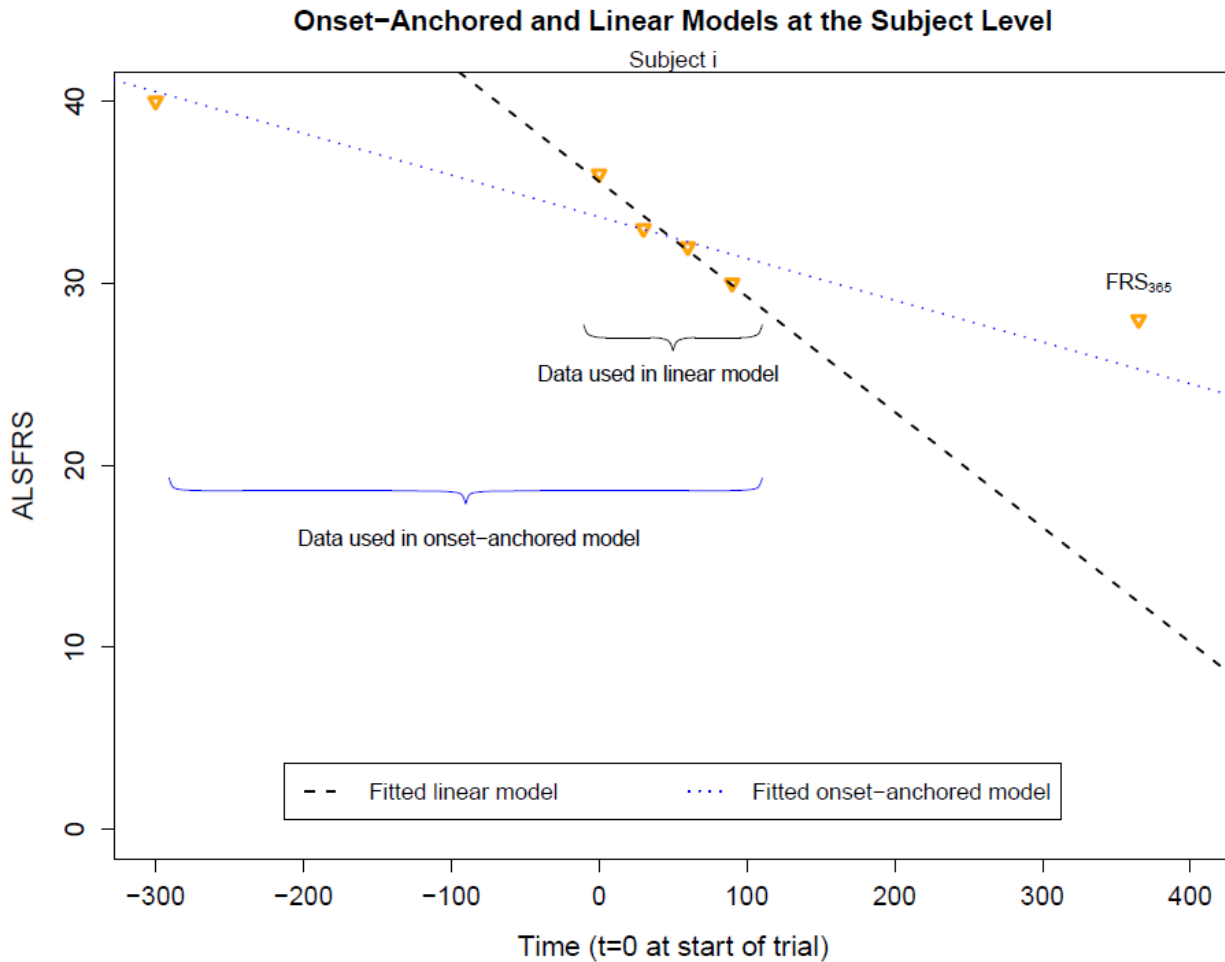


Figure 2.2: Ordinary least-squares estimates for two models: the linear model uses data from zero to three months only, while the onset-anchored model includes an additional artificial data-point. This time point is given as  $(x, y) = (\text{time-point of disease onset-time}, \text{maximal ALSFRS of } 40)$ .

Assuming the maximum possible ALSFRS score at disease onset time was an idea first introduced by Proudfoot et al. (21). They used this assumption to create a slope between the onset anchor and the first observed ALSFRS score that then was used as a predictor for measuring a patient's disease progression. Our onset-anchored model, however, treats this additional artificial data point as an observed value (specifically, a leverage point) in the modelling framework.

Considering the simplicity of this approach, the addition of a non-random leverage point to aid in model prediction is a surprisingly novel technique. This method will, however, result in a biased linear regression model. Recall that the MSE of any prediction is composed of the sum of the square of the bias and the variance of the prediction. In order for this biased model to predict  $FRS_{365}$  well, the reduction in prediction variance needs to dramatically outweigh the increase in prediction bias.

**Covariate selection using the onset-anchored model.** After choosing a “winner” from the three models mentioned above (the onset-anchored model), we wished to determine which clinical features, if any, improved predictive accuracy when used as covariates in the model. Clinical features considered were height, symptom onset time, sex, age, race, individual sub-questions of the ALSFRS, forced vital capacity (FVC, both liters and percent predicted of normal), respiratory rate, weight, Riluzole use (yes/no), and site of onset (bulbar/limb). Many lab measurements are included in PRO-ACT, yet due to their sparse nature, only lab features that were present in at least 90% of the subjects were considered. Albumin has been shown to be associated with ALS survival (22) and was included for analysis even though it was only present in 86% of subjects. The following lab features were considered in our analysis: chloride, serum aspartate aminotransferase (AST), glucose, sodium, blood urea nitrogen (BUN), potassium, bilirubin, alanine transaminase (ALT), creatinine, and albumin.

Many of these features were repeated measures. To use them as covariates, they were truncated to at most three months, then collapsed to slope and intercept (baseline) measures. Specifically, we performed a linear regression on the feature with respect to time, and extracted the ordinary least squares estimates for the slope and intercept. While true baseline data would be preferable over the ordinary least squares intercept estimator, baseline data were frequently not available. Therefore the ordinary least squares intercept estimator was chosen for homogeneity. Collapsing longitudinal predictors has been successfully employed in other ALS predictive models (23, 24), and greatly simplifies the modeling process. All features were normalized using their sample means and variances for ease of analysis and interpretability.

As we were more interested in predictive power, our criteria for feature selection was improvement to the average MSE resulting from predicting  $FRS_{365}$  in repeated cross-validation

analysis (using 90% of the data as the training set and 10% as the validation set), rather than the statistical significance given by a small p-value. Deviance information criterion (DIC) also was considered in assessing whether features improved the model or not.

To assess each covariate we randomly subset the overall data into 300 subjects with non-missing entries. While multiple imputation could be used here, we chose only to use complete cases to drastically reduce computation time as well as eliminate potential convergence problems. We then used cross-validation on this subset of 300 subjects to obtain posterior distributions of the MSE for the onset-anchored model using the covariate and the baseline onset-anchored model for which did not use the covariate. The average difference in MSE between the two models then was computed, and this entire process (starting with subsetting to 300 subjects) was repeated 100 times for each covariate. Specifically, we compared the average difference in MSE of the following two models (for the full model specification, see appendix A1):

Covariate onset-anchored model:

$$ALSFRS_i(t) \sim T_3(b_{0i} + b_{1i}t, \sigma^2)$$

$$b_{0i} \sim N(p_{00} + p_{01}X_i, \sigma_0^2)$$

$$b_{1i} \sim N(p_{10} + p_{11}X_i, \sigma_1^2)$$

Baseline (no covariate) onset-anchored model:

$$ALSFRS_i(t) \sim T_3(b_{0i} + b_{1i}t, \sigma^2)$$

$$b_{0i} \sim N(p_{00}, \sigma_0^2)$$

$$b_{1i} \sim N(p_{10}, \sigma_1^2)$$

Where  $X_i$  is the subject-specific covariate,  $t_i$  is time for subject  $i$ . The slope of subject  $i$  is  $b_{1i}$  which, in the covariate model, is a function depending on  $X_i$ . Similarly,  $b_{0i}$  is the subject-specific intercept. As per hierarchical modelling, we assume priors only for the hyper-parameters  $p_{jk}$  ( $j = 0,1$  and  $k = 0,1$ ). As per the linear model, the following weak priors were assumed:

$$p_{00} \sim N(33, 3^2)$$

$$p_{10} \sim N(-0.025, 0.3^2)$$

Uninformative priors were assigned for the remaining parameters in both models.

### 2.3 Results

We investigated the predictive power of three types of Bayesian hierarchical models: linear, mixture, and onset-anchored. In a Bayesian framework, when cross-validating a model, the resultant MSE has a posterior distribution that takes into account all of the sources of variation within the model. Specifically, these sources of variation include: 1) variation within the model; 2) variation of the posterior parameters; and 3) the variation of the posterior predictive distribution. Therefore, it is important not only to lower the MSE but also to decrease its variance. Of the three models, the onset-anchored model not only had the smallest MSE but also had the MSE with the smallest variance (Figure 2.3). Note that the DIC between the onset-anchored model and the standard linear model cannot be compared, because the additional data point in the onset-anchored model results in a different likelihood.

The MSE for the onset-anchored model not only is smaller in terms of expectation (In Figure 2.3 the means of the MSE for the onset-anchored, mixture model, and linear model were 51.1, 68.5, and 73.7 respectively) but also has the smallest variance. We also considered a mixture model that utilized the additional data point given by the onset anchor. This complex model

performed about as well as the more parsimonious onset-anchored model, that can be seen by their nearly overlapping MSE distributions in Figure 2.3.

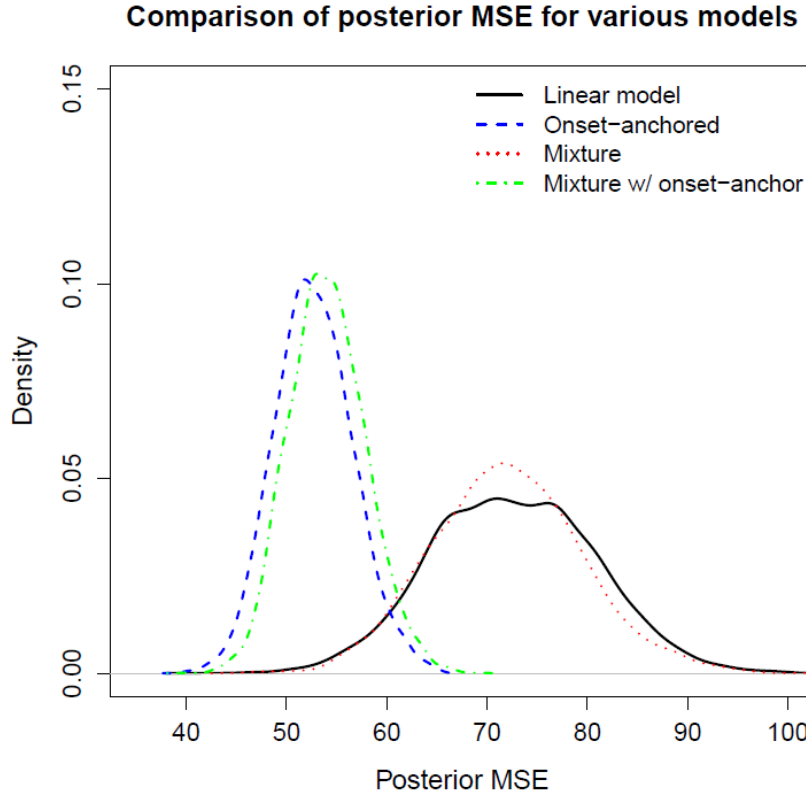


Figure 2.3: Comparison of posterior MSE distribution for four types of hierarchical models: linear, onset-anchored, mixture, and mixture with the additional data-point used in the onset-anchored model. This is from a single cross-validation analysis, but this separation of distributions is typical.

We next attempted to find which covariates or features could consistently improve the MSE of the onset-anchored model, or decrease the DIC. While many clinical and lab predictors had nonzero effects on the posterior slope and intercept (meaning  $p_{11}$  and/ or  $p_{01}$  were nonzero), very few predictors consistently improved the MSE, and among those that did, the improvement to the MSE was very small (Table 2.2). Some variables, such as FVC: Subject Liters (slope) and forced vital capacity (FVC): Percent Normal (slope) reduced DIC (each reduced DIC by about 3.45); however, they did not contribute towards a meaningful improvement in predictive power. Of the 53 covariates tested, only “disease onset time” resulted in an improvement to the MSE



that was on average greater than one percent. This is most likely due disease-onset time giving a slight bias correction to the model. The next best covariates were subject's 3-month slope of FVC (in raw liters) and 3-month slope of the first question from the ALSFRS: Q1, Speech.

<b>Covariate Name</b>	<b>Median % MSE reduction (negative values signify an increase to the MSE)</b>	<b>IQR for % MSE reduction</b>	<b>Mean DIC adjustment (larger values result in larger DIC)</b>
Onset Time	0.0174	0.027	2.8
FVC: Subject Liters (slope)	0.0095	0.02	-3.5
Q1: Speech (slope)	0.0081	0.0187	0.8
Diagnosis Time	0.0059	0.0183	-2.1
Q7: Turning in Bed(slope)	0.0052	0.0182	-3.5
Q8: Walking (slope)	0.0052	0.0254	-2.2
AST (slope)	0.0043	0.0179	-0.3
Q5: Cutting (slope)	0.0043	0.0202	-1.5
Q6: Dressing/Hygiene (slope)	0.0039	0.0223	-1.9
ALT (slope)	0.0034	0.0166	0.1
Q2: Salivation (slope)	0.0032	0.0195	1.7
Q9: Climbing Stairs (slope)	0.003	0.0232	-2
AST (intercept)	0.0028	0.0162	0.8
FVC: Percent Normal (slope)	0.0025	0.0225	-3.4
Race	0.0021	0.0156	-0.1
ALT (intercept)	0.0021	0.0182	-0.7
Bilirubin Total (slope)	0.0019	0.0196	-0.5
Respiratory Rate (intercept)	0.0017	0.0146	0.3
Q2: Salivation (intercept)	0.0013	0.0203	-1.4
Creatinine (intercept)	0.0011	0.0152	-0.7
Age	0.001	0.0185	-2.1
Q1: Speech (intercept)	0.001	0.0224	2.2
Potassium (slope)	0.001	0.0136	-0.9
Onset Site: Bulbar	0.001	0.0171	-0.5
Height	0.0009	0.0169	0.8
Weight (slope)	0.0009	0.0174	-1.7
Sodium (intercept)	0.0008	0.0188	1
Bilirubin Total (intercept)	0.0008	0.0183	0.8
Sex	0.0006	0.0169	-0.4
Q10: Respiratory (slope)	0.0006	0.0153	0.8
Q4: Handwriting (slope)	0.0006	0.0206	-0.7
Weight (intercept)	0.0006	0.0181	-0.5
Q5: Cutting (intercept)	0.0002	0.0224	-4.1

Table 2.2: Median reduction to MSE, in percentage, for covariates which improved the MSE. The inter-quartile range for the percent reduction as well as average difference in DIC is shown as well.

Recall that several of these clinical values have been found to be associated with survival, including Forced Vital Capacity (FVC), age of onset, and site of onset (bulbar or limb, which can help differentiate subtypes of ALS). However, none of these covariates have been consistently useful for modelling ALSFRS progression (9); this is consistent with our findings. Riluzole use, in particular, worsened MSE by a median of 0.09% (see appendix A3 for expanded Table 2.2). Again, this is not surprising as Riluzole has only a weak effect on survival and has not been shown to be consistently associated with decreased disease progression (10, 25).

To appropriately predict the ALSFRS for a given subject at one year from trial onset using data collected up to three months after trial onset, a measure of uncertainty must be reported as well. Since a Bayesian analysis instead was performed, we can obtain 95% credible intervals for each subject's  $FRS_{365}$ . Figures 2.4 and 2.5 give a sample of posterior distributions from a cross-validation for nine randomly-selected subjects'  $FRS_{365}$ , as well as their 95% credible intervals and true  $FRS_{365}$  (the subject's first score at, or after, 365 days). To further demonstrate the improved predictive power of the onset-anchored model, this is done for both the standard linear model (Figure 2.4) as well as the onset-anchored model (Figure 2.5).

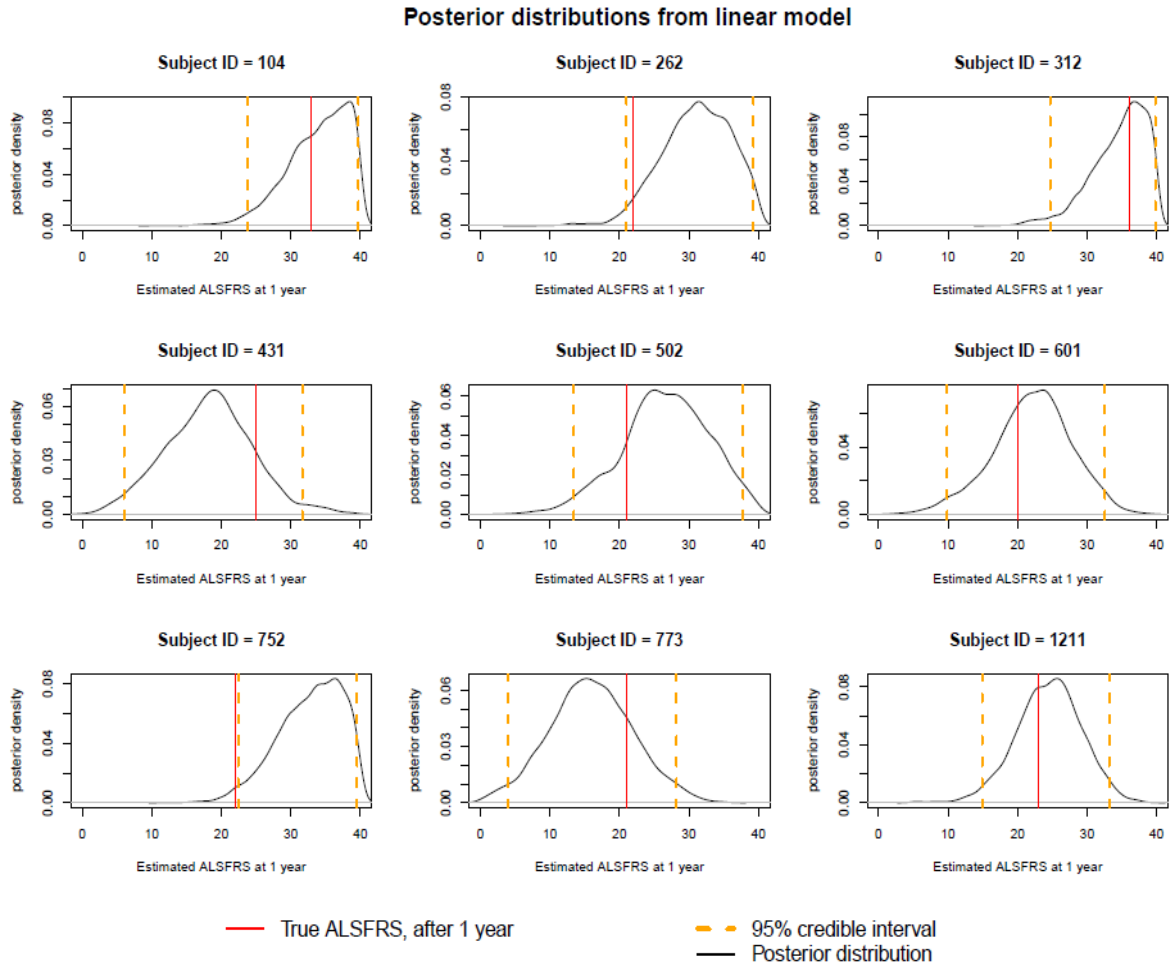


Figure 2.4: Posterior predictive distributions for a random sample of subjects'  $FRS_{365}$  obtained through cross-validation utilizing the standard linear model.

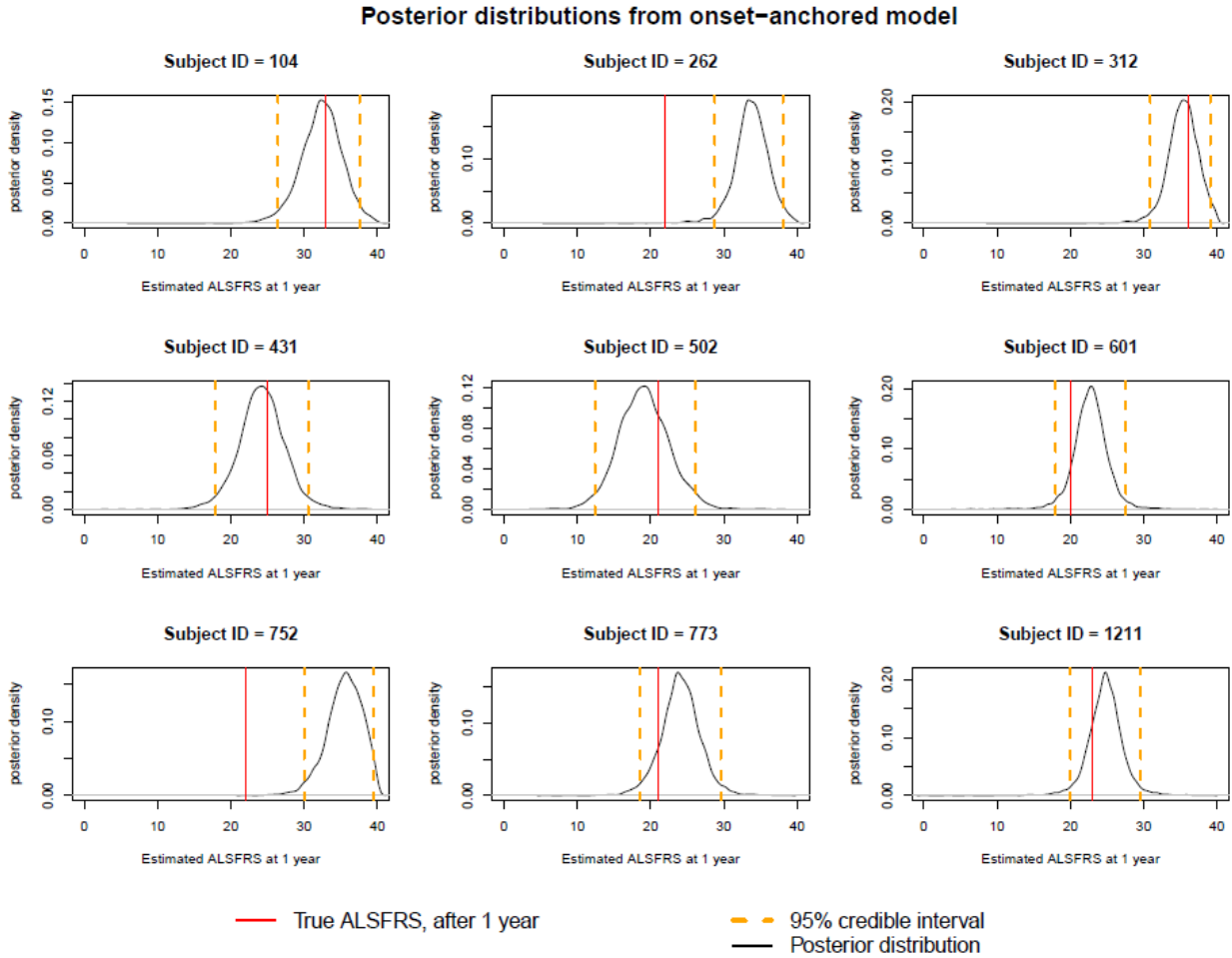


Figure 2.5: Posterior predictive distributions for a random sample of subjects'  $FRS_{365}$  obtained through cross-validation utilizing the onset-anchored model.

It can be noted that the credible intervals for the linear model are very wide, encompassing nearly the full range of the disease. The onset-anchored model drastically reduces the length of these credible intervals. In repeated cross validation, the 95% credible interval contained the true  $FRS_{365}$  for a given subject roughly 73% of the time. As the time of data collection used to make the prediction increases from three months, this prediction becomes more accurate.

The performance of the onset-anchored model vastly is superior to that of the linear model when the length of time for data collection is short. Figure 2.6 shows that the onset-

anchored model, using only baseline data, typically outperforms a linear model using many months of subject data. Figure 2.6 also shows that the onset-anchored model performs well even when the window for data capture is restricted to less than three months, including when only a baseline measurement is available for each subject.

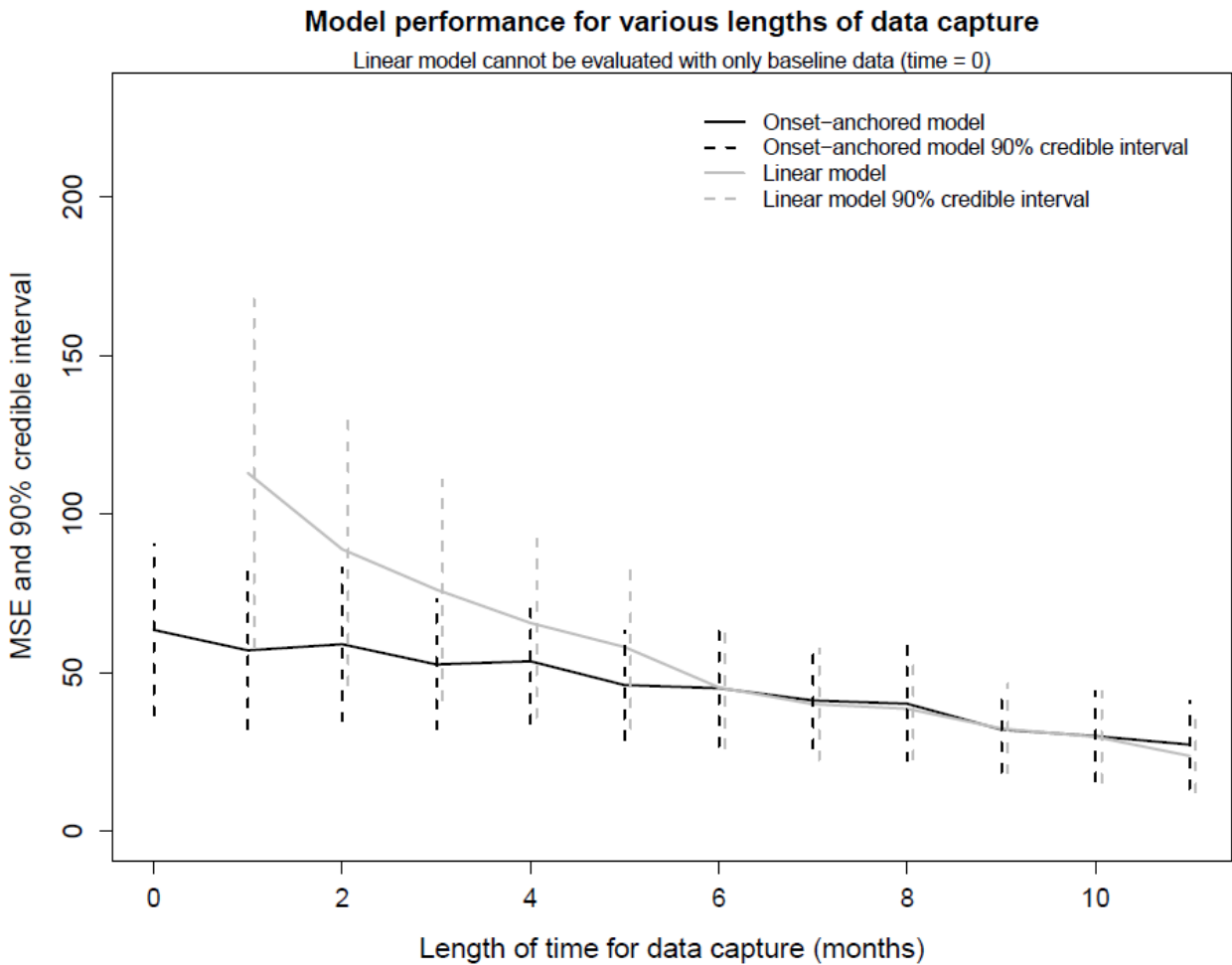


Figure 2.6: The effect of changing the number of months of data used for prediction in both the linear and onset-anchored models. This effect is measured by the MSE (and associated 90% credible interval) resulting from a single cross-validation analysis for both models.

Recall that, while MSE of prediction drastically is reduced when using the onset-anchored model, it is in fact a biased model. The additional data point causes the model typically to underestimate the rate of disease progression, resulting in a higher predicted  $FRS_{365}$  than

observed. Using the onset-anchored model resulted in a prediction bias of, on average, about two (on the ALSFRS scale). For comparison, the linear model typically was unbiased.

Finally, it is typical to measure progression of ALS by the slope of the ALSFRS. An advantage to using the Bayesian framework is that the ALSFRS slope for subject  $i$ , defined previously as  $b_{1i}$ , is specified in the model likelihood, and therefore, has a posterior distribution. Thus, one can then obtain a posterior estimate and credible interval for slope from this distribution. In other words, when using this model one easily can predict slope for a given subject in addition to  $FRS_{365}$ . Examples of the posterior predictive distributions for the ALSFRS slope using three months of data, with 90% credible intervals, is provided in Figure 2.7 for the same nine subjects used in Figures 2.4 and 2.5. As the onset-anchored model performs well even when using only baseline data, subject slopes could be predicted using this model as soon as a baseline ALSFRS score has been established.

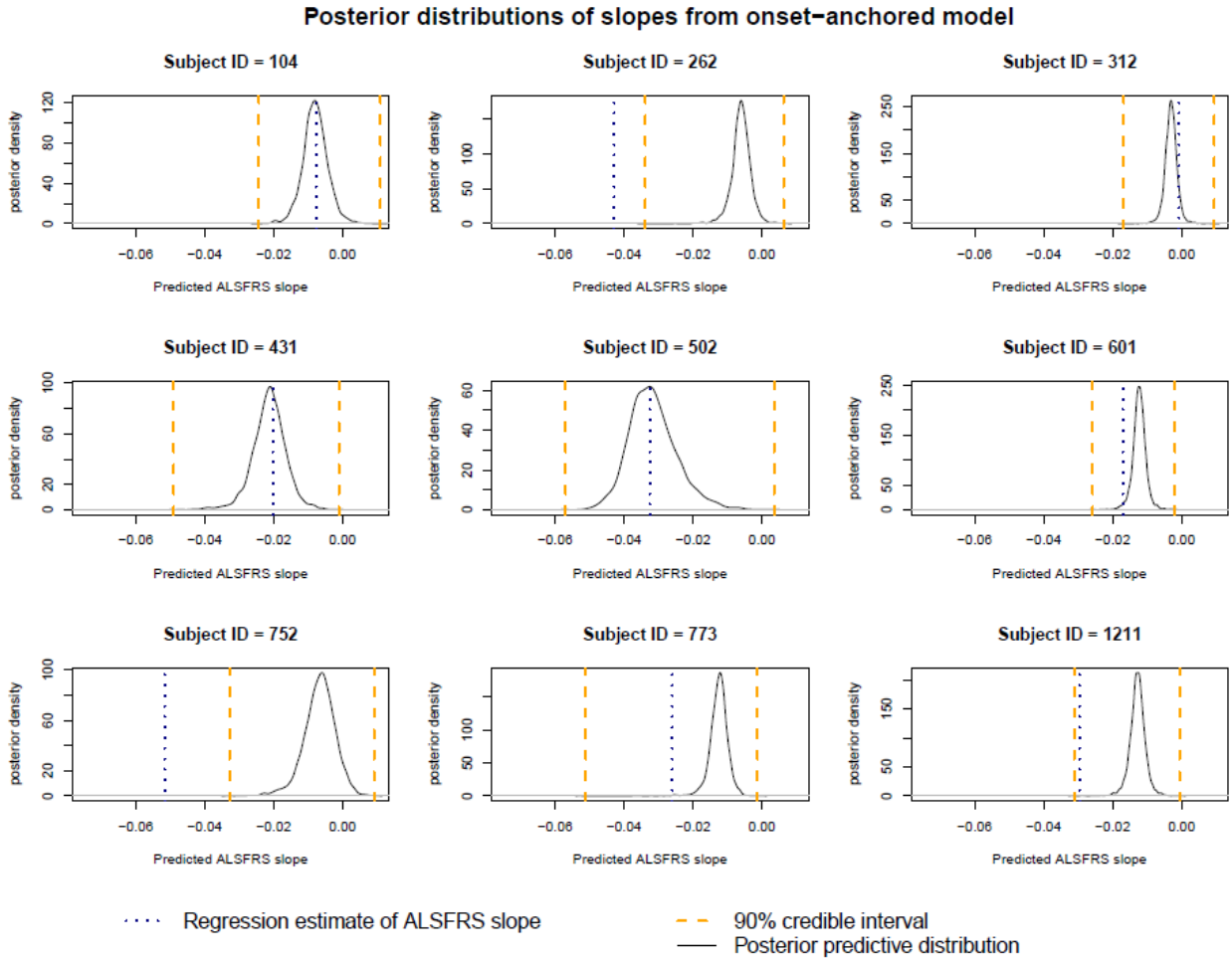


Figure 2.7: Posterior predictive distributions for a random sample of subjects' ALSFRS slope obtained through cross-validation utilizing the onset-anchored model. The regression estimate of the ALSFRS slope (dotted line) was calculated using ordinary least squares on subject's full data.

## 2.4 Discussion

We explored three different Bayesian hierarchical predictive models with the goal of modelling ALS disease progression. These models were linear, mixture, and onset-anchored. The onset-anchored model that uses an additional data point by assuming the maximum ALSFRS score at time of disease onset (e.g. 40), is the best model in terms of predictive accuracy via cross-validation. This is especially noticeable when the window for data capture is very small, such as only using a baseline ALSFRS score.



While linear over the course of a typical clinical trial, progression of the ALSFRS could become curvilinear over long periods of time. This is reinforced further by the fact that it is bounded between zero and 40, and typically is non-increasing. Predictive models that attempt to account for this non-linear progression suffer from a disparity between the number of subject-specific data points and the necessary number of model parameters. We hypothesize that using the onset-anchor helps to “balance” this prediction (see Figure 2.2), while also enabling shrinkage on the slope estimator. The result is a model that has reduced variability of parameter estimates (at the cost of a small increase in bias), which enables a large reduction in overall prediction MSE.

Using three months of subjects’ data, we found that very few clinical features improved prediction as measured by the MSE of repeated cross-validation analysis. Among those features that did consistently improve the MSE, the improvement was rarely more than a one percent reduction. This corroborates findings by Creemers et al. (9) who found the quality of evidence among disease progression prognostic factors to be low at best. The covariate that offered the largest and most consistent improvement to the model’s prediction was disease onset time. As disease onset time is also a key part of the onset-anchor model, this stresses its importance and supports other studies that have shown onset time strongly is associated with disease progression as well as survival (5, 6, 23).

From a practical point of view, a model that only requires time of disease onset and three months of progression data eases both patient and clinician burden by requiring less overall measurements. The Bayesian modeling approach proposed here can help inform the design of adaptive studies, and can be used as an imputation scheme to conduct trials more quickly (26-28). Finally patients with ALS are routinely interested in charting their own progression, as well

as trying interventions that might include treatments for spasticity or pain, or supplements geared towards slowing disease progression. In conjunction with a self-administered ALSFRS, the onset-anchored model then becomes a predictive tool that an ALS patient can use aid them in tracking their disease and assess the utility of self-administered interventions.

While the idea of using an additional data point as used by the onset-anchored model is simple, it is surprisingly novel. Creating biased models to improve predictive MSE is not uncommon, and is used in ideas like fixed-point regression or ridge regression. However, using an artificially created data point and treating it as observed data is something that, to the best of our knowledge, is something that has never been used before. We have found no literature where it theoretically is discussed or practically used. This methodology could be applied to any longitudinal data where the onset time of the process being modelled is known. Other diseases that have bounded rating scales that measure progression, including Parkinson's disease or Huntington's disease, might benefit tremendously from predictions that use an onset-anchor.

One limitation to the current study is that subjects who died before the clinical trial had progressed a full year were not candidates for cross validation, and hence did not directly contribute to the MSE. However, the Bayesian framework allows these subjects to be included in building the model, where their often-increased rates of disease progression contribute to the variability of the model. Specifically, subjects who died prior to one year still contributed towards key model variables, including the distributions of rate of progression, effects of covariates, and variability measures throughout the model. Subjects who died prior to one year also had, on average, a lower predicted  $FRS_{365}$  than subjects who survived past one year. This is expected since a faster progression is associated with lowered survival.

Another limitation is the width of the posterior predictive distributions among individual subjects'  $FRS_{365}$ . These distributions express a combination of variation within the model, variation of the posterior parameters, and variation of the posterior predictive distribution. Due to the heterogeneity of ALS, it is not unexpected that  $FRS_{365}$  can range widely at the individual-patient level. This will remain a limitation of any predictive model until better factors that more strongly are associated with disease progression (rather than survival) are discovered.

The onset-anchored model's inherent bias is another limitation of the model. This is the typical concern with any biased linear model, but in this case we can see that the reduction in the onset-anchored model's MSE is worth the tradeoff. A possible solution might be to investigate a bias-correction term that would utilize disease-onset time as well as the number of days after the start of the trial that is associated with  $FRS_{365}$ .

One final limitation worth pointing out is that disease-onset time, a critical feature of the onset-anchored model, is a problematic variable. This variable typically comes from patient memory, and as a result is subject to recall bias. Proudfoot et al. point out that while this bias exists, using patient-recalled onset time is still a useful predictor for disease progression (21), and this is corroborated by our model.

## 2.5 Conclusions

In this paper we considered the problem of predicting an ALS patient's ALSFRS score at one year, given up to three months of data. Three different Bayesian hierarchical predictive models were considered: linear, mixture, and onset-anchored. The onset-anchored model that leverages an additional artificial data point that assumes the maximum ALSFRS score of 40 at the patient's time of disease onset, is the best model with respect to predictive accuracy under

cross-validation. The onset-anchored model is simple to implement, and potentially is applicable to various other diseases that measure progression by bounded rating scales.

The effect of many covariates (lab values, demographic information, etc.) on these predictions was assessed via repeated cross-validation. The result is that time of disease onset is the only covariate that provides a consistent improvement to predictions, but this is a very small improvement. This highlights the urgent need to develop a better understanding of the mechanism behind ALS progression.

The onset-anchored model has an added benefit over the other models in that it allows predictions to be made as soon as a patient has a baseline ALSFRS score. In other words, as soon as the first ALSFRS measure is taken in a clinical trial, the model can be utilized for endpoint prediction of the ALSFRS. We hope this model can be used by clinicians and statisticians to improve the efficacy of clinical trials and aid in finding treatments for ALS.

## **2.6 References**

1. Armon C, Graves MC, Moses D, Forte DK, Sepulveda L, Darby SM, et al. Linear estimates of disease progression predict survival in patients with amyotrophic lateral sclerosis. *Muscle Nerve*. 2000;23(6):874-82.
2. Magnus T, Beck M, Giess R, Puls I, Naumann M, Toyka KV. Disease progression in amyotrophic lateral sclerosis: predictors of survival. *Muscle Nerve*. 2002;25(5):709-14.
3. Gordon PH, Cheng B, Salachas F, Pradat PF, Bruneteau G, Corcia P, et al. Progression in ALS is not linear but is curvilinear. *J Neurol*. 2010;257(10):1713-7.
4. Ikeda K, Hirayama T, Takazawa T, Kawabe K, Iwasaki Y. Relationships between disease progression and serum levels of lipid, urate, creatinine and ferritin in Japanese patients with amyotrophic lateral sclerosis: a cross-sectional study. *Intern Med*. 2012;51(12):1501-8.

5. Kimura F, Fujimura C, Ishida S, Nakajima H, Furutama D, Uehara H, et al. Progression rate of ALSFRS-R at time of diagnosis predicts survival time in ALS. *Neurology*. 2006;66(2):265-7.
6. Kollewe K, Mauss U, Krampfl K, Petri S, Dengler R, Mohammadi B. ALSFRS-R score and its ratio: a useful predictor for ALS-progression. *J Neurol Sci*. 2008;275(1-2):69-73.
7. Pastula DM, Coffman CJ, Allen KD, Oddone EZ, Kasarskis EJ, Lindquist JH, et al. Factors associated with survival in the National Registry of Veterans with ALS. *Amyotroph Lateral Scler*. 2009;10(5-6):332-8.
8. Zach N, Ennist DL, Taylor AA, Alon H, Sherman A, Kueffner R, et al. Being PRO-ACTive: What can a Clinical Trial Database Reveal About ALS? *Neurotherapeutics*. 2015;12(2):417-23.
9. Creemers H, Grupstra H, Nollet F, van den Berg LH, Beelen A. Prognostic factors for the course of functional status of patients with ALS: a systematic review. *J Neurol*. 2015;262(6):1407-23.
10. Mandrioli J, Biguzzi S, Guidi C, Sette E, Terlizzi E, Ravasio A, et al. Heterogeneity in ALSFRS-R decline and survival: a population-based study in Italy. *Neurol Sci*. 2015;36(12):2243-52.
11. Watanabe H, Atsuta N, Nakamura R, Hirakawa A, Watanabe H, Ito M, et al. Factors affecting longitudinal functional decline and survival in amyotrophic lateral sclerosis patients. *Amyotroph Lateral Scler Frontotemporal Degener*. 2015;16(3-4):230-6.
12. Atassi N, Berry J, Shui A, Zach N, Sherman A, Sinani E, et al. The PRO-ACT database: design, initial analyses, and predictive features. *Neurology*. 2014;83(19):1719-25.
13. Gelman A. Bayesian data analysis. Third edition. ed. Boca Raton: CRC Press; 2014.

14. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.
15. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*. 2000;10(4):325-37.
16. Sturtz S, Ligges U, Gelman A. R2WinBUGS: A Package for Running WinBUGS from R. 2005. 2005;12(3):16.
17. Castrillo-Viguera C, Grasso DL, Simpson E, Shefner J, Cudkowicz ME. Clinical significance in the change of decline in ALSFRS-R. *Amyotroph Lateral Scler*. 2010;11(1-2):178-80.
18. Gomeni R, Fava M, Pooled Resource Open-Access ALSCTC. Amyotrophic lateral sclerosis disease progression model. *Amyotroph Lateral Scler Frontotemporal Degener*. 2014;15(1-2):119-29.
19. Morris CN, Lysy M. Shrinkage Estimation in Multilevel Normal Models. 2012:115-34.
20. Stein C, editor Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; 1956 1956; Berkeley, Calif.: University of California Press.
21. Proudfoot M, Jones A, Talbot K, Al-Chalabi A, Turner MR. The ALSFRS as an outcome measure in therapeutic trials and its relationship to symptom onset. *Amyotroph Lateral Scler Frontotemporal Degener*. 2016:1-12.
22. Chio A, Calvo A, Bovio G, Canosa A, Bertuzzo D, Galmozzi F, et al. Amyotrophic lateral sclerosis outcome measures and the role of albumin and creatinine: a population-based study. *JAMA Neurol*. 2014;71(9):1134-42.

23. Hothorn T, Jung HH. RandomForest4Life: a Random Forest for predicting ALS disease progression. *Amyotroph Lateral Scler Frontotemporal Degener.* 2014;15(5-6):444-52.
24. Kuffner R, Zach N, Norel R, Hawe J, Schoenfeld D, Wang L, et al. Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nat Biotechnol.* 2015;33(1):51-7.
25. Shamshiri H, Fatehi F, Davoudi F, Mir E, Pourmirza B, Abolfazli R, et al. Amyotrophic lateral sclerosis progression: Iran-ALS clinical registry, a multicentre study. *Amyotroph Lateral Scler Frontotemporal Degener.* 2015;16(7-8):506-11.
26. Gajewski BJ, Berry SM, Quintana M, Pasnoor M, Dimachkie M, Herbelin L, et al. Building efficient comparative effectiveness trials through adaptive designs, utility functions, and accrual rate optimization: finding the sweet spot. *Statistics in medicine.* 2015;34(7):1134-49.
27. Rosenblum M, Lubert B, Thompson RE, Hanley D. Group sequential designs with prospectively planned rules for subpopulation enrichment. *Statistics in medicine.* 2016.
28. Shan G, Wilding GE, Hutson AD, Gerstenberger S. Optimal adaptive two-stage designs for early phase II clinical trials. *Statistics in medicine.* 2016;35(8):1257-66.

## Chapter 3: Automated Data Extraction of Electronic Medical Records to Model ALS

### Survival and Progression

#### 3.0 Abstract

**Objective:** To assess the feasibility of using automated capture of Electronic Medical Record (EMR) data for future amyotrophic lateral sclerosis (ALS) clinical research by comparing a University of Kansas Medical Center (KUMC) ALS data to an established ALS clinical trial population provided by The Pooled Resource Open-Access ALS Clinical Trials database (PRO-ACT).

**Methods:** We used an Informatics for Integrating Biology and the Bedside search discovery tool to identify and extract 219 ALS patients from the KUMC EMR, which were compared to 1300 placebo-arm subjects from PRO-ACT. Bayesian hierarchical modelling provided estimates of disease progression as measured by the ALS Functional Rating Scale. Two different Cox proportional hazards models were used to investigate the effect of KUMC membership on survival.

**Results:** KUMC patients were typically older at disease onset, were more likely to have bulbar onset, had slower ALS progression, and had improved survival rates versus subjects in PRO-ACT. We found the following to be associated with improved survival: younger age, more time from ALS onset to baseline measure, limb (versus bulbar) onset, and higher baseline BMI, forced vital capacity, and ALS Functional Rating Scale scores.

**Conclusion:** We show the feasibility of using automated data extraction from the EMR to model and track disease progression in ALS. Differences in disease progression and survival in the KUMC patients compared to the PRO-ACT placebo controls highlight the need to better



understand disease variability in the clinical setting, and to refine the inclusion criteria in ALS trials.

### **3.1 Introduction**

Amyotrophic Lateral Sclerosis (ALS) is a fatal neuro-degenerative disease. While over 50 clinical trials have been conducted over the last two decades, none have been successful save riluzole and edaravone, that at best offer modest improvements in survival or function (1). While many studies may have failed because the drugs were ineffective, a recurring theme in ALS is underpowered trials with indeterminate results. Two major hurdles to conducting ALS trials are the rarity of ALS (3.9 in every 100,000 people in the U.S. (2)) and the disease's heterogeneity.

Obtaining historic controls or creating predictive models from Electronic Medical Record (EMR) data would be one way to improve inexpensively the power and efficiency of ALS trials (3, 4). The wide use of EMRs, and the development automated abstraction of de-identified data, create opportunities to: 1) better understand ALS disease progression and determinates of survival in the clinical setting; 2) use clinical data to enrich existing placebo-arm data sets to improve the power of trials; and 3) leverage this electronic infrastructure to run clinical trials – including EMR-based screening, randomization, and data collection. For these approaches to be valid, we need to demonstrate the feasibility of modeling disease progression using EMR-captured clinical data.

Here we compared statistical models built with EMR-captured patient data from our ALS clinic at the University of Kansas Medical Center (KUMC) to a typical ALS clinical trial population provided by The Pooled Resource Open-Access ALS Clinical Trials database (PRO-ACT) (5). KUMC's EMR is provided by Epic.

### 3.2 Methods

**Study design.** We conducted a retrospective chart review of patients at the KUMC ALS Clinic. To do this efficiently, we used the KUMC Healthcare Enterprise Repository for Ontological Narration (HERON), powered by Informatics for Integrating Biology and the Bedside (i2b2), a discovery tool that allows searches of de-identified EMR data (6, 7). This allowed us to extract ALS patient data from the EMR for analysis, with records dating from summer 2012 to early 2017. This data include information on patient demographics, disease progression, and survival. Survival data captured by HERON comes from both the EMR and from the Social Security Death Index (8). This data then were verified and augmented by a manual chart review (Table 3.1).

	<b>Information extracted from HERON</b>	<b>Information extracted from chart review</b>
<b>Demographic data</b>	Subject age*, ethnicity*, race*, gender*, marital status, religion, date of birth	Date of disease onset*, date of diagnosis*, site of disease onset*, ALS family history
<b>Longitudinal data</b>	ALSFRRS-R and its sub scores*, BMI*, FVC (raw and percent-predicted)*	Negative inspiratory force (NIF)
<b>Medication history</b>		Riluzole*, Nuedexta, Vitamin use
<b>Other</b>	Death status*, date of death*	Non-invasive positive pressure ventilation (NIPPV) status, date of NIPPV, NIPPV type

Table 3.1: Specific data extracted from the KUMC EMR by HERON, and data added from manual chart review. Asterisk (\*) denotes the data was also available in PRO-ACT.

**Study populations.**

*University of Kansas Medical Center (KUMC) ALS Clinic data.* The ALS clinic at KUMC serves roughly four state regions across the Midwest (Kansas, Missouri, Oklahoma, Arkansas). At each visit, patient data collected by the clinician are entered in the EMR (EPIC EMR system, Epic Systems Corporation, 2015). Using HERON, we extracted data from patients who met the criteria of having at least one ALS Functional Rating Scale – Revised (ALSFRS-R) score, that measures ALS progression. Patients' ALSFRS-R is recorded in-clinic by a research coordinator and entered into a document flow sheet. Only patients seen in the ALS specialty clinic with a known diagnosis of motor neuron disease have ALSFRS-R scores in the EMR. Initially, the KUMC data from HERON covered 357 subjects and 2,848 individual clinical visits. We eliminated patient encounters not pertaining to ALS, death, or respiratory progression. We further eliminated patients whose first and last ALSFRS-R scores occurred less than 90 days apart or had a diagnosis other than ALS (such as primary lateral sclerosis). This resulted in 219 unique subjects and 960 clinical records.

***Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) data.*** In 2011, Prize4Life, in collaboration with the Northeast ALS Consortium, and with funding from the ALS Therapy Alliance, formed the PRO-ACT Consortium. The data available in the PRO-ACT Database have been volunteered by PRO-ACT Consortium members (<https://nctu.partners.org/PRO-ACT/>) (5). We used PRO-ACT records prior to December 1, 2015, which had 4,838 unique subjects with progression information (ALSFRS-R scores, or, more commonly, ALSFRS scores). Only subjects that were assigned to the placebo arm of their trials were considered so as to not contaminate our statistical models with unknown effects from various therapeutics. This resulted in 1,301 subjects and 11,773 patient records. Demographic information, including a comparison between the PRO-ACT data and the KUMC data, is given in the results section.

**Outcomes.** The specific variables we obtained pertaining to demographic, disease progression, and survival are provided in Table 3.1. Disease progression of ALS typically is measured by the ALSFRS-R that consists of twelve physician-administered questions. Each question is on a 0 – 4 scale, with the overall score of 48 representing normal function (9). The ALSFRS-R has become the standard for measuring disease progression, and easily can be converted to the previous version, the ALSFRS. The ALSFRS, typically reported in older clinical trials, is the first 10 questions of the ALSFRS-R (meaning a score of 40 represents normal function on the ALSFRS).

For this study, we utilized the ALSFRS (rather than the ALSFRS-R) in order to maximize the data available from the PRO-ACT database. Note that in ALS trials, the ALSFRS typically declines by one ALSFRS point per month (10).

## Statistical methods.

**Disease progression.** We used a Bayesian hierarchical linear model (with error terms following a  $T$  distribution with three degrees of freedom) to model disease progression of the KUMC and PRO-ACT cohorts. While a Random Forest approach, such as performed by Taylor et al. (11), is accurate in predicting future ALSFRS-R scores, a hierarchical model allows us to easily estimate cohort-level disease progression rates. Quantile-quantile plots showed a  $T$  distribution with three degrees of freedom was more appropriate than the normal distribution for both the KUMC and PRO-ACT error terms.

A linear model was selected due to the known typical linear decline of the ALSFRS (12, 13) that was confirmed via visual inspection of the KUMC data. Some authors have disputed this in favor of a quadratic model (14, 15), but we found the quadratic term to be insignificant via deviance information criterion (DIC). A major advantage to utilizing a linear model is the convenience of using the ALSFRS slope parameter as an easily-interpreted measure of disease progression.

The model is described as  $ALS_i(t) \sim T_3(a_i + b_i t, \sigma^2)$ , restricted to  $ALS_i \in [0, 40]$ , with patient-specific intercept  $a_i \sim N(p_0, \sigma_0^2)$  and slope  $b_i \sim N(p_1, \sigma_1^2)$ . This means the posterior distribution of  $p_1$  models the cohort-level disease progression. Weak priors were assumed from  $p_0$  and  $p_1$  (Supplement Appendix A1), and  $b_i$  was restricted to be non-positive. This approach is the Bayesian analogue to the linear mixed effects model.

The predictive power of this model can be improved by using an “onset-anchor”. This is equivalent to augmenting each subject’s data with an additional fixed data-point given by assuming maximal ALSFRS score at time of their disease onset (16). This causes a slight bias in the model at the cost of drastically reduced variability. Since the mean squared error (MSE) of

prediction is equivalent to the sum of the variability (variance) and the square of the bias, a large reduction in variability at the cost of a slight increase in bias can result in models with superior predictions. We also adjusted for time of disease onset, which acts as a bias-correction term.

***Survival.*** We used disease-onset time as the baseline survival time for all patients rather than time of first clinical visit. This is necessary to create a common baseline time for measuring survival, as there is no reason to suggest that patients' first clinical visit would be distributed equivalently across the two data sets. Survival then was modelled utilizing the Cox proportional hazards model. Note that time of disease onset typically is dependent on patients recalling when they first began having symptoms, which can be error-prone (16).

An advantage to the Cox model is that we can test models that use several covariates at once. This allows us to consider the KUMC membership (versus PRO-ACT) effect on survival, after adjusting for other significant variables (17). To obtain such a model, we used a “backwards with forward looks” approach, only considering records without missing data. Covariates of interest were those that previously have been shown to be independently associated with survival in ALS. These include rate of ALS progression (15, 18-20), forced vital capacity (FVC) (21, 22), age at symptom onset (23, 24), site of symptom onset (13, 20, 23, 25), and body mass index (BMI) (22, 23). Specifically, we considered the following baseline covariates: sex, riluzole use, years from onset to baseline, age, FVC (percent predicted), BMI, ALSFRS score, diagnostic delay (months between disease onset and diagnosis of ALS), and site of disease onset. Certain data were omitted for extreme sparsity (such as negative inspiratory force) or for not being available in both data sets (such as marital status). All analyses were done using R (version 3.2.4) (26) and OpenBUGS (version 3.2.3 rev 1012) (27).

### 3.3 Results

**Comparison of datasets.** Race and gender were similar across the two datasets (Table 3.2). Both were roughly 60% male and predominantly white, agreeing with reports published by the United States Center for Disease Control (CDC) (2). The typical site of disease onset was in the limbs, although limb onset was more prevalent in PRO-ACT. Both had roughly the same percentage of observed deaths; however, the follow-up time for the KUMC patients was on average longer than those in PRO-ACT. KUMC patients had roughly half the average total number of clinical visits versus PRO-ACT subjects.

	<b>KUMC</b>	<b>PRO-ACT</b>
<b>Number of patients</b>	219	1300
<b>Percent Female/Male</b>	44/ 56	38/ 62
<b>Percent Caucasian / Non-Caucasian</b>	90/ 10	94/ 6
<b>Percent Limb onset / Bulbar / Other</b>	66/ 30/ 4	79/ 20/ 1
<b>Percent using riluzole Yes / No / Unknown</b>	72/ 28/ 0	46/ 28/ 26
<b>Percent deceased during follow up</b>	35	32.1
<b>Mean Follow-up time days (SD)</b>	436.7 (309.5)	294.7 (164)
<b>Mean Age at disease onset (SD)</b>	60.1 (11.4)	53.7 (12.0)
<b>Mean Number of months from onset to diagnosis (SD)</b>	23.08 (33.45)	11.5 (8.8)
<b>Mean First recorded ALSFRS score (SD)</b>	28.2 (6.3)	29.63 (5.8)
<b>Mean Years from onset to first clinical record (SD)</b>	2.5 (3.2)	1.8 (1.3)
<b>Mean FVC percent predicted at baseline (SD)</b>	69 (21)	78 (21)
<b>Mean BMI at baseline (SD)</b>	27.37 (5.7)	25.6 (4.7)
<b>Mean number of visits (SD)</b>	4.38 (2.27)	9.06 (4.4)

Table 3.2: Comparison of KUMC and PRO-ACT information. Baseline is defined as the time of a patient's first recorded ALSFRS score.

Other differences between the two cohorts included: 1) average age at disease onset, that was older in the KUMC dataset; 2) subjects at KUMC had lower initial ALSFRS scores and baseline FVC (percent normal); 3) the number of months from disease onset to diagnosis (diagnostic delay) was longer in KUMC than PRO-ACT, with increased variability – a result of 14 KUMC patients with diagnostic delay greater than five years, whereas PRO-ACT had only one.

**Analysis of disease progression.** Cohort-level disease progression was modelled by the posterior distribution of  $p_1$  (see methods), that portrays the possible ALSFRS slopes of a given cohort (Figure 3.1). We observed that PRO-ACT subjects had more rapid disease progression (typical difference of 0.17 ALSFRS points per month, or 2.00 points per year). The 90% credible interval for KUMC disease progression was (-0.89, -0.76), compared to (-1.02, -0.95) for PRO-ACT. The only variable that consistently improved the predictions given by the model, via cross-validation, was time of disease onset.

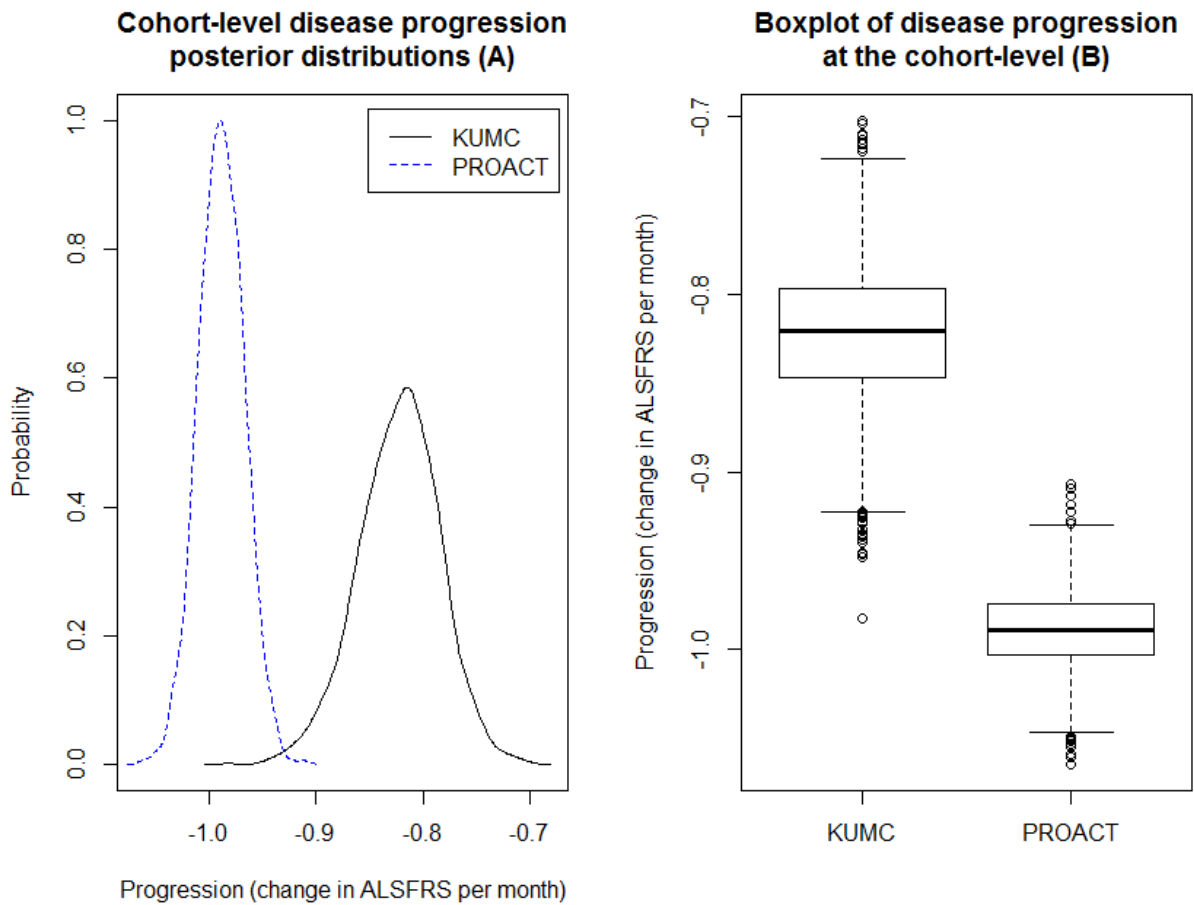


Figure 3.1: Posterior estimates of cohort-level disease progression for KUMC and PRO-ACT patients, given both as kernel densities (A) and boxplots (B). As the progression is measured as the change in ALSFRS score per month, a more negative value corresponds to a more rapid rate of disease progression.



**Analysis of survival.** Median survival time from disease onset was 2.8 years for KUMC, compared to 2.2 years for PRO-ACT. The raw Kaplan-Meier survival curves (unadjusted for other covariates) obtained from the left-truncated, right-censored model is given in Figure 3.2. Note that the KUMC survival curve is never below the PRO-ACT survival curve, which allowed us to proceed with the Cox model.

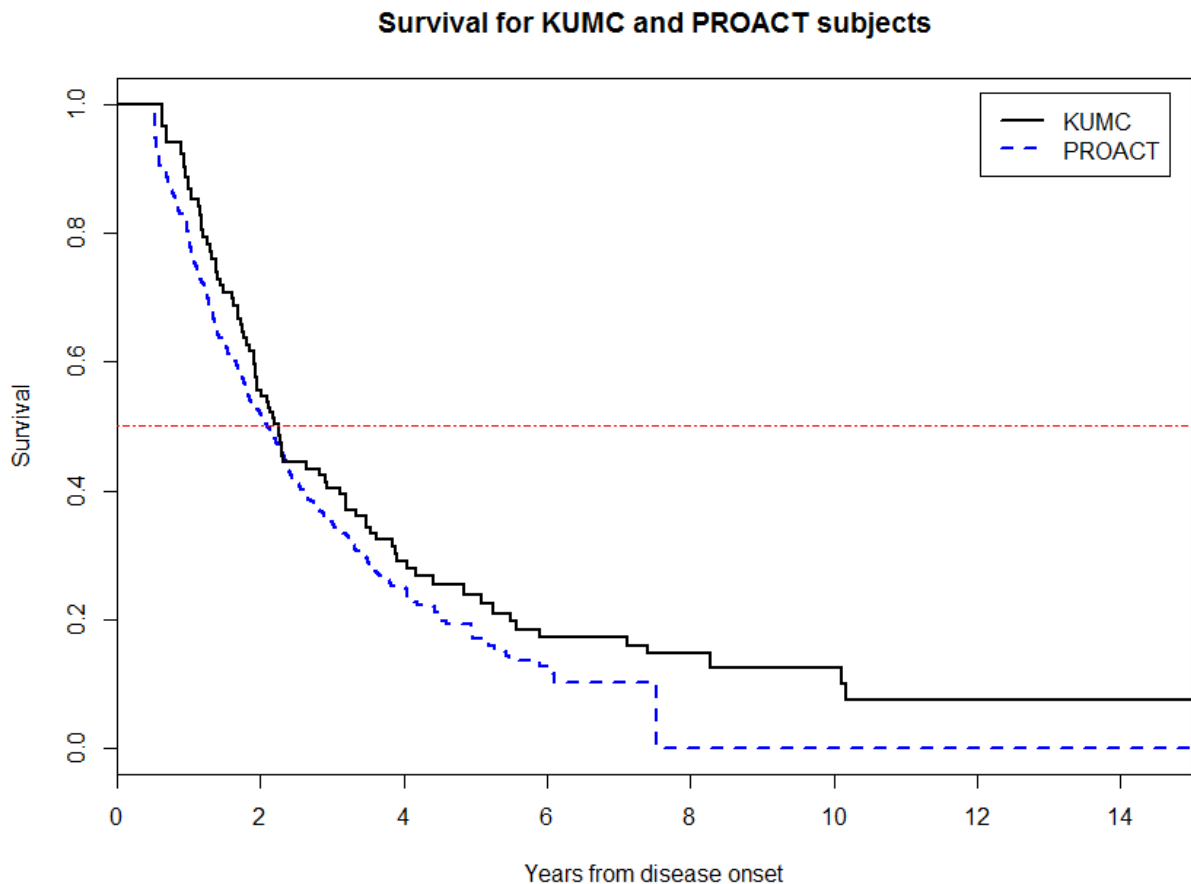


Figure 3.2: Kaplan-Meier survival curves for KUMC and PRO-ACT subjects. Where the horizontal line intersects the curve represents the median survival time, equivalent to where the y-axis is 0.5.

The final Cox model for the two datasets included KUMC/PRO-ACT membership, riluzole use, years from onset to baseline, age at baseline, FVC at baseline (percent normal), BMI at baseline, and ALSFRS total at baseline. However, our modelling approach used complete

cases only (meaning subjects only were included if they had no missing data among the considered covariates), so this model includes only 78 of the 219 KUMC patients and 619 of 1,300 PRO-ACT subjects. This is due to a large amount of missing baseline BMI and FVC information. The effects of each covariate, along with standard errors, are given in Table 3.3 under “Model 1”. KUMC membership, longer time from onset to baseline, younger age, and larger baseline FVC, BMI, and ALSFRS resulted in an improved survival time. Riluzole use was associated with a relatively strong decrease in survival.

<b>Model 1 (Complete cases only)</b>			
<b>Covariate</b>	<b>Fitted coefficient</b>	<b>Standard error</b>	<b>P-value</b>
KUMC membership	-2.0136	0.2804	<0.0001
Riluzole use	1.4540	0.1450	<0.0001
Years from onset to baseline	-1.1557	0.1794	<0.0001
Age (baseline)	0.0424	0.0061	<0.0001
FVC (baseline, percent normal)	-0.0229	0.0047	<0.0001
BMI (baseline)	-0.0599	0.0151	0.0001
ALSFRS total (baseline)	-0.0785	0.0116	<0.0001
<b>Model 2 (All subjects)</b>			
<b>Covariate</b>	<b>Fitted coefficient</b>	<b>Standard error</b>	<b>P-value</b>
KUMC membership	-0.7411	0.1418	<0.0001
Years from onset to baseline	-0.7369	0.1177	<0.0001
Age (baseline)	0.0383	0.0042	<0.0001
ALSFRS total (baseline)	-0.1071	0.0071	<0.0001
Site of onset (limb onset)	-0.3330	0.1020	0.0011

Table 3.3: Coefficient values for the best fitting Cox survival models. Model 1 is found by only considering to complete cases (no missing data). This model may be misleading because it utilizes a small subset of the KUMC and PRO-ACT subjects. Model 2 utilizes all KUMC and PRO-ACT subjects but disregards baseline BMI and FVC as potential covariates.

Eliminating baseline BMI and FVC in model selection (each are associated independently with survival after controlling for KUMC/PRO-ACT membership) results in a Cox model that utilizes all 219 subjects from KUMC and 1,300 from PRO-ACT. In this model, we again see that KUMC membership, longer time from onset to baseline, younger age, and

larger baseline ALSFRS resulted in an improved survival time (Table 3.3). Additionally, limb onset was associated with improved survival time. Note that the individual effect of group membership has been reduced greatly in this model; being in the KUMC cohort still results in improved survival, but the effect is not as before.

### **3.4 Discussion**

A barrier to EMR data abstraction is the development of tools to automate the data extraction and the creation of a common data dictionary to pool data across multiple sites. As part of KUMC's participation in the Patient Centered Outcomes Research Institute (PCORI) sponsored Greater Plains Collaborative, a consortium of nine academic centers across the Midwest, we had access to HERON, a data abstraction tool using i2b2 technology. This enabled us easily to extract patient data from the KUMC EMR, which has fields vital for measuring disease progression and survival in ALS. The KUMC ALS clinic routinely has been collecting the ALSFRS and FVC to track progression and manage care in the clinic. We demonstrated the feasibility of: (1) electronically extracting this data from KUMC's EMR, and (2) successfully using this data to create models of KUMC patient progression and survival.

As a positive control for the approach, our survival analyses corroborate previously found survival results. Our models showed strong evidence that the number of years from ALS onset to first clinical visit, age, and first recorded ALSFRS score are associated with survival (significant in all considered survival models). There was supporting evidence that site of disease onset, baseline FVC, and baseline BMI also are associated with survival (significant in one of the two considered survival models). All explanatory variables had the expected beneficial or detrimental effect on survival (18-25). We were able to extract several key survival variables from the EMR and confirm that they were significant while modelling survival.

As more clinics use their EMR systems to capture important measures accurately for ALS, such as the ALSFRS, it will become possible to leverage large, complete, EMR-based, ALS datasets. These datasets will create opportunities to design new varieties of trials and hasten ALS treatment discovery. One simple improvement to ALS trials would come from using EMR data to augment trial data, as either a placebo arm or as historical controls. In order for this to be feasible, we need a better understanding of the variability and progression of patients in the clinic setting. Despite seeing the same clinical predictors of survival as other studies, we found ALS data derived from the clinic were different in both progression and survival when compared to participants recruited for clinical trials as represented by PRO-ACT. Our statistical models revealed that the subjects in PRO-ACT had faster disease progressions than KUMC patients. Even though KUMC patients were more likely to be older and have bulbar onset (commonly associated with shorter survival time (13, 20, 23, 25)), KUMC patients still had improved survival versus PRO-ACT subjects, even after adjusting for significant survival predictors.

This raises a flag of caution; these preliminary results suggest that the rich source of patient data from the EMR may not correspond to what we have been seeing in ALS trials, even after adjusting for baseline differences between the clinic and PRO-ACT. Before using this approach larger data sets for modeling progression and survival in the clinic setting may be required. It also raises the question, how generalizable are the results of ALS trials to the general population?

One major reason for the differences between these two datasets are the trial inclusion/exclusion criteria. In most of the PRO-ACT trials, these criteria were designed to select patients early in the disease process (symptom onset within 2-3 years of study entry, baseline FVC > 60-75% predicted), with the hope of increasing the likelihood for patients to survive to

the end of the trial. Recent advances in our understanding of ALS progression has revealed this likely is due to the heterogeneous patient group selected by inclusion/exclusion criteria that includes both slow and rapid progressing patients (28).

A simple solution would be to modify the inclusion/exclusion criteria of trials to be more representative of the general population; however, the resulting increased patient variability would require larger studies. Yet this is readily solvable: with a more general trial population, we would be free to use the EMR to augment greatly the control population for these trials. Networks such as the Greater Plains Collaborative could provide placebo or standard-of-care arms in a variety of designs that could make such large-scale studies possible.

In our analysis, riluzole use was associated with decreased survival. However, the datasets in this paper do not represent a true unbiased population, as patients were not *randomized* to riluzole. Patients included here self-selected for riluzole use. As a result, this association most likely was due to confounding or some unknown bias; perhaps only the very sick or rapidly deteriorating person decided to take riluzole.

A limitation to this study is that the KUMC data comes from one ALS clinic. Follow up would require validation of the approach in a multi-center cohort. However, this study serves as proof of principle that data automatically abstracted from the EMR can be used for modeling disease progression and survival in ALS. An additional limitation is data missing from the EMR. In order to use EMR-collected data for more sophisticated analysis, clinicians and clinic staff will need to ensure complete data entry that includes concomitant medications, participation in clinical trials, vital statistics, and agreement on a common minimal ALS data set. For example, information on patients' usage of PEG tubes may be of interest, but was not captured in the EMR.

In the future, data abstracted from the EMR may fill multiple roles in ALS research, including using computable phenotypes for study screening (3), modeling the effect of repurposed drugs and supplements, the effects of drugs taken for other medical conditions on progression or survival, and running pragmatic studies to determine the best symptomatic treatments or timing of interventions (BIPAP, PEG) on clinical outcomes (4). EMR data could also be used as a concurrent standard of care control arm, and could be useful in creating large-sample predictive models (29). This study serves as proof of concept for the approach of using automatically abstracted data from the EMR to model progression and survival.

### 3.5 References

1. Goyal NA, Mozaffar T. Experimental trials in amyotrophic lateral sclerosis: a review of recently completed, ongoing and planned trials using existing and novel drugs. Expert opinion on investigational drugs. 2014;23(11):1541-51.
2. Mehta P, Kaye W, Bryan L, Larson T, Copeland T, Wu J, et al. Prevalence of Amyotrophic Lateral Sclerosis - United States, 2012-2013. Morbidity and mortality weekly report Surveillance summaries. 2016;65(8):1-12.
3. Beaulieu-Jones BK, Greene CS, Pooled Resource Open-Access ALSCTC. Semi-supervised learning of the electronic health record for phenotype stratification. Journal of biomedical informatics. 2016;64:168-78.
4. Pearson JF, Brownstein CA, Brownstein JS. Potential for electronic health records and online social networking to redefine medical research. Clinical chemistry. 2011;57(2):196-204.
5. Atassi N, Berry J, Shui A, Zach N, Sherman A, Sinani E, et al. The PRO-ACT database: design, initial analyses, and predictive features. Neurology. 2014;83(19):1719-25.

6. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association : JAMIA. 2010;17(2):124-30.
7. Waitman LR, Warren JJ, Manos EL, Connolly DW. Expressing observations from electronic medical record flowsheets in an i2b2 based clinical data repository to support research and quality improvement. AMIA Annual Symposium proceedings AMIA Symposium. 2011;2011:1454-63.
8. Security S. The Full Death Master File: SSA; 2017 [cited 2016]. Available from: [https://www.ssa.gov/dataexchange/request\\_dmf.html](https://www.ssa.gov/dataexchange/request_dmf.html).
9. Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). J Neurol Sci. 1999;169(1-2):13-21.
10. Castrillo-Viguera C, Grasso DL, Simpson E, Shefner J, Cudkowicz ME. Clinical significance in the change of decline in ALSFRS-R. Amyotroph Lateral Scler. 2010;11(1-2):178-80.
11. Taylor AA, Fournier C, Polak M, Wang L, Zach N, Keymer M, et al. Predicting disease progression in amyotrophic lateral sclerosis. Annals of clinical and translational neurology. 2016;3(11):866-75.
12. Armon C, Graves MC, Moses D, Forte DK, Sepulveda L, Darby SM, et al. Linear estimates of disease progression predict survival in patients with amyotrophic lateral sclerosis. Muscle Nerve. 2000;23(6):874-82.
13. Magnus T, Beck M, Giess R, Puls I, Naumann M, Toyka KV. Disease progression in amyotrophic lateral sclerosis: predictors of survival. Muscle Nerve. 2002;25(5):709-14.

14. Gomeni R, Fava M, Pooled Resource Open-Access ALSCTC. Amyotrophic lateral sclerosis disease progression model. *Amyotroph Lateral Scler Frontotemporal Degener.* 2014;15(1-2):119-29.
15. Gordon PH, Cheng B, Salachas F, Pradat PF, Bruneteau G, Corcia P, et al. Progression in ALS is not linear but is curvilinear. *J Neurol.* 2010;257(10):1713-7.
16. Proudfoot M, Jones A, Talbot K, Al-Chalabi A, Turner MR. The ALSFRS as an outcome measure in therapeutic trials and its relationship to symptom onset. *Amyotroph Lateral Scler Frontotemporal Degener.* 2016:1-12.
17. Klein JP, Moeschberger ML. *Survival analysis : techniques for censored and truncated data.* 2nd ed. New York: Springer; 2003. xv, 536 p. p.
18. Kimura F, Fujimura C, Ishida S, Nakajima H, Furutama D, Uehara H, et al. Progression rate of ALSFRS-R at time of diagnosis predicts survival time in ALS. *Neurology.* 2006;66(2):265-7.
19. Kollewe K, Mauss U, Krampfl K, Petri S, Dengler R, Mohammadi B. ALSFRS-R score and its ratio: a useful predictor for ALS-progression. *J Neurol Sci.* 2008;275(1-2):69-73.
20. Elamin M, Bede P, Montuschi A, Pender N, Chio A, Hardiman O. Predicting prognosis in amyotrophic lateral sclerosis: a simple algorithm. *J Neurol.* 2015;262(6):1447-54.
21. Czaplinski A, Yen AA, Appel SH. Forced vital capacity (FVC) as an indicator of survival and disease progression in an ALS clinic population. *Journal of neurology, neurosurgery, and psychiatry.* 2006;77(3):390-2.
22. Paganoni S, Deng J, Jaffa M, Cudkowicz ME, Wills AM. Body mass index, not dyslipidemia, is an independent predictor of survival in amyotrophic lateral sclerosis. *Muscle Nerve.* 2011;44(1):20-4.



23. Zach N, Ennist DL, Taylor AA, Alon H, Sherman A, Kueffner R, et al. Being PRO-ACTive: What can a Clinical Trial Database Reveal About ALS? *Neurotherapeutics*. 2015;12(2):417-23.
24. Testa D, Lovati R, Ferrarini M, Salmoiraghi F, Filippini G. Survival of 793 patients with amyotrophic lateral sclerosis diagnosed over a 28-year period. *Amyotrophic lateral sclerosis and other motor neuron disorders : official publication of the World Federation of Neurology, Research Group on Motor Neuron Diseases*. 2004;5(4):208-12.
25. Pastula DM, Coffman CJ, Allen KD, Oddone EZ, Kasarskis EJ, Lindquist JH, et al. Factors associated with survival in the National Registry of Veterans with ALS. *Amyotroph Lateral Scler*. 2009;10(5-6):332-8.
26. R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2016.
27. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. *Statistics in medicine*. 2009;28(25):3049-67.
28. Chio A, Canosa A, Gallo S, Cammarosano S, Moglia C, Fuda G, et al. ALS clinical trials: do enrolled patients accurately represent the ALS population? *Neurology*. 2011;77(15):1432-7.
29. Paxton C, Niculescu-Mizil A, Saria S. Developing predictive models using electronic medical records: challenges and pitfalls. *AMIA Annual Symposium proceedings AMIA Symposium*. 2013;2013:1109-15.

## Summary and Future Directions

Predicting disease progression for patients with ALS could be useful for patients, doctors, and clinical trial investigators. Patients could self-track and compare their progression rates to those predicted in order to assess the effectiveness of self-administered therapeutics. Doctors could use predictions to help their patients to assess how their disease will progress and patients plan their future actions accordingly. Clinical trial investigators could benefit immensely from predictive models as well. These models could be used in adaptive designs to make better-informed decisions on randomization schemes and early stopping rules. They could also serve as mechanisms for performing imputation.

Two challenges when predicting ALS disease progression are (1): the sparsity of data, which is due to the rareness of ALS, and (2): the heterogeneity of progression rates among patients. Because the disease tends to decline in a linear fashion, linear models are appropriate for modelling ALS progression. Linear models enjoy a plethora of advantages over other models with respect to interpretability, but suffer from large prediction variance when a subject has few observed data points. We sought to improve this prediction variance by exploiting knowledge of when a patient first began having ALS symptoms. We could use this knowledge to create an additional data point (an anchor) that could then be used to reduce the variability of the linear predictions.

We demonstrated that an anchor could be used theoretically, or in practice, when modeling several ALS patients' disease progressions under a standard OLS framework (chapter one). We also considered more complex models, such as the Bayes hierarchical linear model, and found that using an anchor resulted in improvements with respect to model prediction (chapter two). We sought to build the best predictive model for ALS disease progression that we

could: this entailed analyzing various potential covariates for improvements to prediction, and rigorously testing these models via repeated cross-validation. Finally, we demonstrated that these models could be used for both ALS data from the University of Kansas Medical Center's ALS clinic (automatically extracted via the EMR), as well as from the PRO-ACT database of ALS clinical trials (chapter three).

Future works include the comparison of our model to more recently popularized ALS predictive models that utilize random forest approaches. We ultimately would like to expand the work in chapter three to multiple clinics (rather than solely the clinic at KUMC), so as to get a more representative sample of the general ALS population. Finally, one of the key purposes to developing ALS predictive models is to use them to design efficient clinical trials which could hasten the development of new ALS therapeutics. Therefore it would be prudent to develop the framework for a Bayesian adaptive clinical trial that would incorporate these predictive models to test new ALS treatments. This framework then could be compared to frequentist designs via simulation, incorporating historical ALS data to demonstrate the advantage of using predictive models in ALS trial design. It would be our hope that this design could be used in the future to assist in developing of new ALS therapeutics.

## Appendix

### Appendix A1: Full model description

Let  $X_i$  be a covariate and  $ALSFRS_i(t)$  the ALSFRS score for subject  $i$  at time  $t$ . In this,  $\sigma^2$  is the precision parameter of the non-standardized  $T$  distribution (3 degrees of freedom), which is defined as the inverse of the variance. The linear and onset-anchored hierarchical models are then defined as follows:

$$ALSFRS_i(t) \sim T_3(b_{0i} + b_{1i}t, \sigma^2); ALSFRS_i(t) \in [0, 40]$$

$$b_{0i} \sim N(p_{00} + p_{01}X_i, \sigma_0^2); b_{0i} \in [0, 40]$$

$$b_{1i} \sim N(p_{10} + p_{11}X_i, \sigma_1^2); b_{1i} \in (-\infty, 0]$$

$$p_{00} \sim N(33, 0.111)$$

$$p_{01} \sim N(0, 0.0001)$$

$$p_{10} \sim N(-0.025, 11)$$

$$p_{11} \sim N(0, 0.0001)$$

$$\sigma^2 \sim \Gamma(0.001, 0.001)$$

$$\sigma_0^2 \sim \Gamma(0.001, 0.001)$$

$$\sigma_1^2 \sim \Gamma(0.001, 0.001)$$

## Appendix A2: BUGS code (R2 OpenBUGS format)

```

for (i in 1:N) #N is number of subjects.
{
  for (j in 1:J[i]) #J is a vector containing the number of entries for each subject.
  {
    ALSmu[Jtot[i]+j] <- b0[i]+b1[i]*delta[Jtot[i]+j]
    ALS[Jtot[i]+j] ~ dt(ALSmu[Jtot[i]+j], err, 3)%_%I(0,40)

  }
  b1_mean[i] <- p10 + p11*covariate[Jtot[i]+1]
  b0_mean[i] <- p00 + p01*covariate[Jtot[i]+1]

  b1[i]~dnorm(b1_mean[i],err1) %_%I(,0)
  b0[i]~dnorm(b0_mean[i],err0) %_%I(0,40)

  #predictive draws
  predmu[i] <- b0[i]+b1[i]*365
  pred365[i] ~ dt(predmu[i], err, 3) %_%I(0,40)

}

##priors
p00 ~ dnorm(33, 0.111)
p01 ~ dnorm(0, 0.00001)
p10 ~ dnorm(-0.025, 11)
p11 ~ dnorm(0, 0.00001)

err~dgamma(0.001, 0.001)
err1~dgamma(0.001, 0.001)
err0~dgamma(0.001, 0.001)

```

## Appendix A3: Full table the effects of covariates on prediction under the onset anchored

### model

Covariate Name	Median % MSE reduction (negative values signify an increase to the MSE)	IQR for % MSE reduction	Mean DIC adjustment (larger values result in larger DIC)
Onset Time	0.0174	0.027	2.8
FVC: Subject Liters (slope)	0.0095	0.02	-3.5
Q1: Speech (slope)	0.0081	0.0187	0.8
Diagnosis Time	0.0059	0.0183	-2.1
Q7: Turning in Bed(slope)	0.0052	0.0182	-3.5

Q8: Walking (slope)	0.0052	0.0254	-2.2
AST (slope)	0.0043	0.0179	-0.3
Q5: Cutting (slope)	0.0043	0.0202	-1.5
Q6: Dressing/Hygiene (slope)	0.0039	0.0223	-1.9
ALT (slope)	0.0034	0.0166	0.1
Q2: Salivation (slope)	0.0032	0.0195	1.7
Q9: Climbing Stairs (slope)	0.003	0.0232	-2
AST (intercept)	0.0028	0.0162	0.8
FVC: Percent Normal (slope)	0.0025	0.0225	-3.4
Race	0.0021	0.0156	-0.1
ALT (intercept)	0.0021	0.0182	-0.7
Bilirubin Total (slope)	0.0019	0.0196	-0.5
Respiratory Rate (intercept)	0.0017	0.0146	0.3
Q2: Salivation (intercept)	0.0013	0.0203	-1.4
Creatinine (intercept)	0.0011	0.0152	-0.7
Age	0.001	0.0185	-2.1
Q1: Speech (intercept)	0.001	0.0224	2.2
Potassium (slope)	0.001	0.0136	-0.9
Onset Site: Bulbar	0.001	0.0171	-0.5
Height	0.0009	0.0169	0.8
Weight (slope)	0.0009	0.0174	-1.7
Sodium (intercept)	0.0008	0.0188	1
Bilirubin Total (intercept)	0.0008	0.0183	0.8
Sex	0.0006	0.0169	-0.4
Q10: Respiratory (slope)	0.0006	0.0153	0.8
Q4: Handwriting (slope)	0.0006	0.0206	-0.7
Weight (intercept)	0.0006	0.0181	-0.5
Q5: Cutting (intercept)	0.0002	0.0224	-4.1
Q3: Swallowing (slope)	-0.0003	0.0214	-0.3
Q10: Respiratory (intercept)	-0.0005	0.0189	-0.4
Albumin (slope)	-0.0006	0.0183	0.2
Albumin (intercept)	-0.0006	0.0184	-0.6
Creatinine (slope)	-0.0007	0.0192	1
Chloride (slope)	-0.0008	0.0201	-0.9
Used Riluzole	-0.0009	0.0196	-0.6
FVC: Percent Normal (intercept)	-0.0012	0.0177	-0.3
Blood Urea Nitrogen (intercept)	-0.0012	0.017	-1.5
Potassium (intercept)	-0.0012	0.015	0.6
Q3: Swallowing (intercept)	-0.0015	0.016	-0.5

FVC: Subject Raw Liters (intercept)	-0.0016	0.0216	-1.9
Sodium (slope)	-0.0018	0.0187	-0.2
Chloride (intercept)	-0.0021	0.0188	0.2
BloodUrea Nitrogen (slope)	-0.0022	0.02	-1.2
Respiratory Rate (slope)	-0.0027	0.0174	-1.4
Q7: Turning in Bed (intercept)	-0.0039	0.029	-5.9
Q4: Handwriting (intercept)	-0.0043	0.0209	-4.2
Q6: Dressing/Hygiene (intercept)	-0.005	0.0238	-7.8
Q8: Walking (intercept)	-0.0123	0.0226	-1.7
Q9: Climbing Stairs (intercept)	-0.0139	0.0232	-2.8

## **Appendix B1: My Experience at the KUMC ALS Clinic**

While researching ALS at KUMC, I had the opportunity to visit the KUMC ALS clinic and meet actual patients who suffered from the disease. A commonly known struggle with studying ALS is the heterogeneity of the disease, and this was very apparent through the data. However, I truly experienced the between-subject variability first hand while visiting the clinic. Of the three different patients I met, each of them had drastically different manifestations of ALS symptoms.

The first, a middle-aged woman with an incredible optimism, couldn't move her legs and was bound to a wheelchair. Her left arm was beginning to show weakness as well. The next, an elderly man, had complete control of his limbs, but was unable to speak or control his mouth, resulting in a constant trickle of drool coming from his mouth. The man spoke by writing on a notepad he kept with him. The final patient I met had not had the disease for long, and still had complete control of her limbs and head; however she had a pervasive muscle weakness throughout her body that was worsening over time.

From my research that ultimately came down to working with spreadsheets and electronic data, I knew that ALS was a devastating disease. Seeing it in firsthand in these people, many of whom came with their spouses, family members, or children, was tragic. Yet, the amount of positivity and hope displayed by the patients was overwhelming. They were excited and interested in my research, and were very much intrigued by the idea of being able to predict how their disease would progress over the next year. Knowing that my work may be able to improve the lives of these patients, in even a small way, gave me a tremendous amount of motivation. Truly, it is imperative that as biostatisticians that we remember that the trials we design and analyze are: (1) being performed on real people, with all the complexities that entails, (2) have



the potential to benefit thousands, if not millions, and (3) give patients hope; hope of a new treatment, hope of a cure, or hope of an improved quality of life. Keeping all of this in mind, while not getting lost in the numbers and calculations, can be difficult, but remembering these facts is of paramount importance. My visit to the ALS clinic was a great reminder of that.

## **Appendix B2: ALSFRS and ALSFRS-R Questionnaire**

The response to each question, from top to bottom, is worth four, three, two, one, or zero points. The ALSFRS is found by summing the scores from items one through ten. The ALSFRS-R is found by summing the scores from items one through twelve.

### **1. Speech**

- 4: Normal speech processes
- 3: Detectable speech disturbance
- 2: Intelligible with repeating
- 1: Speech combined with nonvocal communication
- 0: Loss of useful speech

### **2. Salivation**

- 4: Normal
- 3: Slight but definite excess of saliva in mouth; may have nighttime drooling
- 2: Moderately excessive saliva; may have minimal drooling
- 1: Marked excess of saliva with some drooling
- 0: Marked drooling; requires constant tissue or handkerchief

### **3. Swallowing**

- 4: Normal eating habits
- 3: Early eating problems-occasional choking
- 2: Dietary consistency changes
- 1: Needs supplemental tube feeding
- 0: NPO (exclusively parenteral or enteral feeding)

### **4. Handwriting**

4: Normal

3: Slow or sloppy; all words are legible

2: Not all words are legible

1: Able to grip pen but unable to write

0: Unable to grip pen

5. Cutting food with gastrostomy

4: Normal

3: Somewhat slow and clumsy, but no help needed

2: Can cut most foods, although clumsy and slow; some help needed

1: Food must be cut by someone, but can still feed slowly

0: Needs to be fed

6. Dressing and hygiene

4: Normal function

3: Independent and complete self-care with effort or decreased efficiency

2: Intermittent assistance or substitute methods

1: Needs attendant for self-care

0: Total dependence

7. Turning in bed

4: Normal

3: Somewhat slow and clumsy, but no help needed

2: Can turn alone or adjust sheets, but with great difficulty

1: Can initiate, but not turn or adjust sheets alone

0: Helpless

## 8. Walking

- 4: Normal
- 3: Early ambulation difficulties
- 2: Walks with assistance
- 1: Non-ambulatory functional movement only
- 0: No purposeful leg movement

## 9. Climbing stairs

- 4: Normal
- 3: Slow
- 2: Mild unsteadiness or fatigue
- 1: Needs assistance
- 0: Cannot do

## 10. Dyspnea

- 4: None
- 3: Occurs when walking
- 2: Occurs with one or more of the following: eating, bathing, dressing (ADL)
- 1: Occurs at rest, difficulty breathing when either sitting or lying
- 0: Significant difficulty, considering using mechanical respiratory support

## 11. Orthopnea

- 4: None
- 3: Some difficulty sleeping at night due to shortness of breath. Does not routinely use more than two pillows
- 2: Needs extra pillow in order to sleep (more than two)

1: Can only sleep sitting up

0: Unable to sleep

12. Respiratory insufficiency

4: None

3: Intermittent use of BiPAP

2: Continuous use of BiPAP

1: Continuous use of BiPAP during the night and day

0: Invasive mechanical ventilation by intubation or tracheostomy

### **Appendix B3: Funding and Support**

This research was made possible with the help of Dr. Byron Gajewski, Dr. Jeffrey Statland, and Dr. Jianghua He, as well as support from the Mabel A. Woodyard Fellowships in Neurodegenerative Disorders and the Roofe Fellowship in Neuroscience Research.

The Pooled Resource Open-Access ALS Clinical Trials Consortium, data used in the preparation of this article, were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. As such, the following organizations and individuals within the PRO-ACT Consortium contributed to the design and implementation of the PRO-ACT Database and/or provided data, but did not participate in the analysis of the data or the writing of this report: Neurological Clinical Research Institute, MGH; Northeast ALS Consortium; Novartis; Prize4Life; Regeneron Pharmaceuticals, Inc.; Sanofi; Teva Pharmaceutical Industries, Ltd.

Additionally, this work was supported by a Clinical Translation Science Award (CTSA) grant from National Center for Research Resources (NCRR) and National Center for Advancing Translational Sciences (NCATS) awarded to the University of Kansas Medical Center for Frontiers: The Heartland Institute for Clinical and Translational Research # UL1TR000001 (formerly #UL1RR033179). The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH, NCRR, or NCATS. Doctor Jeffrey Statland's work on this project was supported by a fellowship grant from the NCATS / Clinical Research in ALS and Related Disorders for Therapeutic Development Consortium that was awarded to the University of Miami (U54NS092091).